CHAPTER 2

ESTIMATION AND PROJECTION OF LIFETIME EARNINGS

ABSTRACT

This chapter describes the estimation and prediction of age-earnings profiles for American men and women born between 1931 and 1960. The estimates are obtained using lifetime earnings records maintained by the Social Security Administration. These data have been combined with demographic information for the same individuals collected in the Survey of Income and Program Participation. The estimates show a substantial rise in lifetime earnings inequality over time and in average lifetime wages earned by American women as compared with men. In addition they show that Baby Boom workers born immediately after the Second World War are likely to enjoy higher average wages relative to economy-wide average earnings than generations born before or after them. The advantage of this cohort over earlier generations is in large measure attributable to major increases in educational attainment. The advantage over later generations is partly due to a small advantage in educational attainment, especially among men, but is primarily due to the very poor job market conditions facing younger members of the Baby Boom generation when they entered the labor force. These adverse conditions persisted for nearly two decades. Under the assumptions of the earnings model estimated here, this early disadvantage will permanently reduce relative lifetime earnings of workers in later Baby Boom cohorts in comparison with the relative earnings enjoyed by the oldest members of the Baby Boom.

I. INTRODUCTION

In order to make forecasts of future Social Security outlays, the future distribution of Social Security pensions and other retirement income, and future impacts on benefits and retirement incomes of changes in the Social Security program, it is necessary to make a prediction of the future level and distribution of labor earnings. Workers' wages and self-employment income determine their eligibility for Social Security benefits and affect the level of benefits and other retirement income to which they will become entitled.

This chapter describes a method for estimating the earnings function that generates typical patterns of career earnings. It is based on a straightforward application of an individual effects statistical model, applied to a rich source of panel data on lifetime earnings. The chapter is organized as follows. The next section describes the estimation problem and statistical approach taken in this project, and the following section describes the data, the empirical estimates, and our methods for making earnings projections based on these estimates. The last section examines some statistical properties of the forecasts.

II. DESCRIPTION OF ESTIMATION PROCEDURES

The profile of annual earned income over the lifetime has a characteristic hump-shaped pattern for typical Americans. Initial earnings are low, reflecting workers' initially modest levels of job tenure, skill, and experience. Earnings rise over time, often in an erratic pattern, as workers accumulate human capital and find jobs that offer wages reflecting the workers' greater skill and job experience. Earnings then fall, either abruptly, as a result of worker retirement or disability, or more gradually, as a result of declining work hours, employer discrimination, or the eroding value of a worker's skills

The characteristic pattern of lifetime earnings profiles is displayed in Figures 2-1 and 2-2, which show the cross-sectional pattern of earned income among women and men, respectively. The higher line in each figure shows the age profile of earnings among all workers who had positive earned incomes in 1996. The profile is estimated as a quadratic function of age using Census Bureau tabulations of average earnings within broad age categories (age 18-24, 25-34, 35-44, and so on). For both women and men the age pattern of earned income, conditional on having positive earnings, shows a rapid rise from ages 22 through 40, slower earnings growth for workers in their 40s, and earnings declines beginning sometime after age 50.

The lower and heavier line in the two figures shows the lifetime profile of average earnings calculated using information from *all* potential workers, including those who do not work. This line shows lower average earnings at each age, especially among women, but it reveals the same characteristic pattern of rapidly rising income when workers are in their 20s and 30s and declining earnings when they are in their 50s and 60s. The estimated peak of expected earnings occurs at an earlier age when people with zero earnings are included in the tabulations. This is because the unconditional earnings profile also incorporates the effects of labor force withdrawal of workers who become disabled or who retire. Since disability and early retirement become more common as workers reach their 50s, the fall in unconditional earnings begins at a younger age.

The lines in the two figures clearly do not represent the earnings experiences of *each* U.S. worker. Instead they reflect the experiences in a single year of all workers when their experiences are averaged together. The cross-sectional pattern of earnings differs widely for workers with different characteristics. The figures show that the patterns for women and men differ noticeably, for example. In comparison with workers who have limited education, workers

Figure 2-1
Age-Earnings Profile of U.S. Women, Including and Excluding Zero Earnings

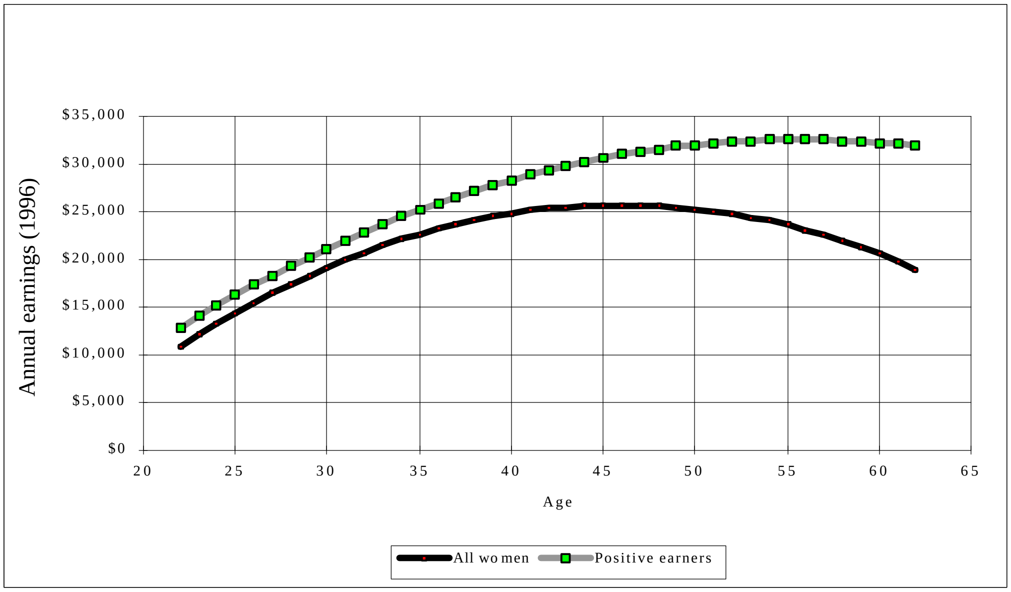
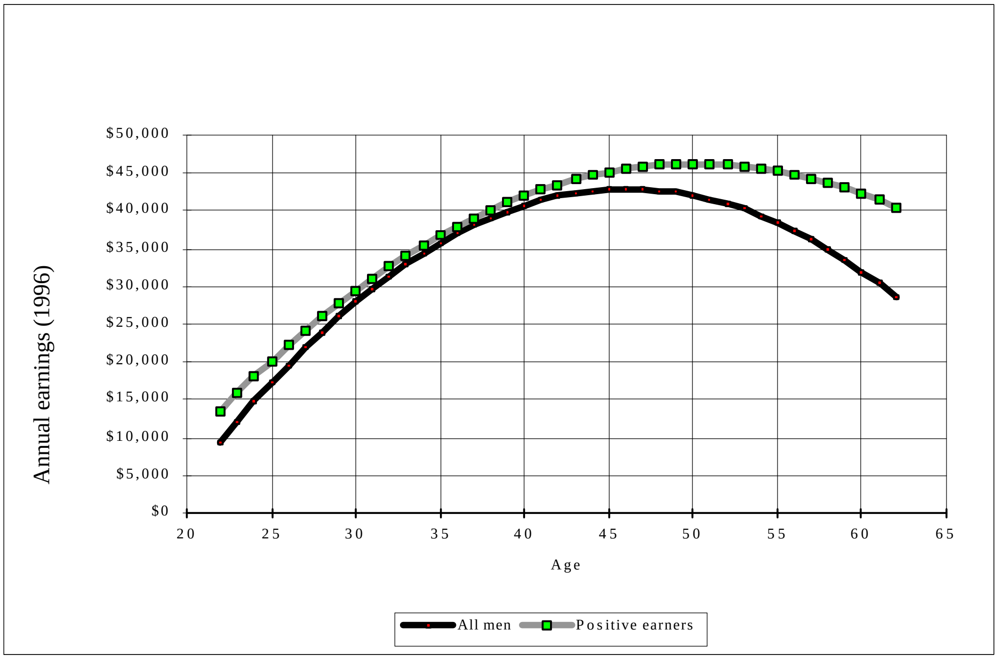


Figure 2-2
Age-Earnings Profile of U.S. Men, Including and Excluding Zero Earnings



with more schooling show a characteristic pattern of steeper earnings growth in their 20s and 30s, and their earnings typically reach a lifetime peak at a later age. The age profile of earnings has not remained fixed over the past few decades. In the 1960s, the cross-sectional age pattern of earnings showed smaller earnings differences between 25-year-old and 45-year-old workers. In other words, the age profile of earnings is now more steeply sloped than it was in the past. Finally, individual workers differ widely from one another. Even among workers with identical observable characteristics, including age, educational attainment, occupational attachment, and job tenure, there are enormous variations in annual earnings and in the pattern of year-to-year earnings change.

1. Basic Specification

To make a forecast of future earnings for workers who have only partially completed their careers, it is necessary to make credible predictions about the structure of future age-earnings profiles. We adopted a simple specification of the basic relation between workers' ages and the change in their earnings. Individual-level earnings is treated as a step-function of age. In particular,

$$y_{it} = \mu_i + f(\text{Age}) + \epsilon_{it}, \qquad (1)$$

where

$$f(\text{Age}) = \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \ldots + \beta_T A_T \text{, and}$$

A_1 = 1 if Age is less than 25,
= 0, otherwise;

A_2 = 1 if Age is between 25 and 29,
= 0, otherwise;

A_3 = 1 if Age is between 30 and 34,
= 0, otherwise;

A_4 = 1 if Age is between 35 and 39,
= 0, otherwise; ***[This category is omitted in the estimation.]***

A_5 = 1 if Age is between 40 and 44,
= 0, otherwise;

A_6 = 1 if Age is between 45 and 49,
= 0, otherwise;

A_7 = 1 if Age is between 50 and 54,
= 0, otherwise;

A_8 = 1 if Age is between 55 and 57,
= 0, otherwise;

A_9 = 1 if Age is between 58 and 59,
= 0, otherwise;

A_{10} = 1 if Age is between 60 and 61,
= 0, otherwise;

A_{11} = 1 if Age is 62,
= 0, otherwise;

A_{12} = 1 if Age is between 63 and 64,

= 0, otherwise;
A_{13} = 1 if Age is 65,
= 0, otherwise;
A_{14} = 1 if Age is 66 or more,
= 0, otherwise.

Ignoring μ_i and ϵ_{it}, this specification implies that earnings rise by varying amounts, β_A, at each of the age breaks specified in the function f(Age). This specification is obviously far more flexible than the quadratic function used to estimate the cross-sectional age-earnings profiles in Figures 2-1 and 2-2.

Economists have scant basis for predicting the future trend of economy-wide average earnings. This trend will obviously have a crucial influence on the earnings profile of workers who are currently young or middle-aged. Rather than estimate the trend in economy-wide earnings directly, we estimate the relationship between workers' *relative* earnings and their age. Relative earnings in this study is defined as the ratio of a worker's earnings in a given year to the economy-wide average covered wage estimated by the Social Security Administration. Thus, the coefficients β_A in equation (1) refer to the change in a worker's relative earnings at each of the age breaks in the age-earnings function, f(Age). If economy-wide average earnings climb rapidly, the β's will be associated with steep growth in actual earnings during the phase in a worker's career when his or her relative earnings are climbing. If economy-wide real wages are stagnant or declining, the β's will be associated with very modest or even shrinking annual earnings.

As noted above, the pattern of career earnings differs across population groups. Earnings profiles differ between men and women and among workers with differing levels of educational attainment. In this study, we estimated separate earnings functions for men and women, who in turn are divided into five educational groups: those who did not complete high school; those with a high school diploma but no schooling beyond high school; those with one to three years of college education; those with a college diploma; and those with at least one year of education beyond college. Workers can of course be divided into even narrower categories, for example, by race, occupational attachment, marital status, and geographic region. In order to keep the estimation and projection simple, we decided not to examine career earnings profiles in these narrower groups. Several of them, including occupation and marital status, can change over a worker's career. Since we observe these time-varying variables only up through the time an individual is last interviewed, we cannot reliably predict how these variables will change over the remainder of the worker's career. For this reason, we do not think it makes sense to include them at this stage in the estimation model.

We estimated the earnings equation under a fixed-effect specification. That is, we assume that each person in a given sub-population differs from other workers in his or her peer group by a fixed average amount. This individual-specific difference persists over a worker's entire career and is captured by the error term μ_i in equation 1 above. Under the assumptions of the fixed-effect model, we cannot obtain estimates of coefficients of variables that do not change over time for a single observation. The effects of these variables are all captured by the person-specific individual effect. Thus, we do not obtain coefficient estimates in the earnings regressions of the effects of a person's race or birth cohort, because these variables do not change over time for people in the sample. (If analysts want to know the average effects of these variables, they can calculate the average value of the estimated fixed effects of respondents with the relevant characteristics.)

The coefficients of the age terms, β_A ,are essentially determined by the average observed change in relative earnings as workers move up from one age category to the next. For example, the coefficient β_3 shows the average difference in earnings between ages 30-34 and the omitted age category, ages 35-39. This is determined by an estimate of the average gain in relative earnings that persons actually experienced between ages 30-34, on the one hand, and ages 35-39, on the other. This kind of estimate can only be obtained with longitudinal information for a sample of workers. (It is *not* an estimate of the average difference in earnings between people who are 30-34 and people who are 35-39 in a given year.)

For estimates based on this model to be valid, it must be the case that future *relative* earnings increases will mirror the pattern observed during the period covered by the estimation sample. Suppose the sample consists of people born between 1931 and 1960, and earnings are observed for the period from 1981 to 1990. The oldest people in the sample are between 50 and 60 years old during the estimation period. From the experiences of these people we can form estimates of the average increase or decline in earnings that takes place between ages 50-54, 55-57, and 58-59. Under the assumptions of the model, the relative earnings gains or losses experienced by this cohort will be duplicated by later cohorts when they reach ages 50-54, 55-57, and 58-59. Of course, the actual average earnings of younger cohorts will differ from those of the older cohort. The model offers two possible explanations for the difference. First, if economy-wide earnings grow faster when the younger cohorts are between 50 and 60, their actual earnings will grow faster (or decline more slowly) than was the case for the older cohort. Second, the average value of the individual specific error term, μ_i , may differ between the two cohorts, although the difference between two large birth cohorts will probably be small.

2. Employment Patterns

The specification defined by equation 1 represents a single-equation model of the earnings generation process. We emphasize that this approach does not adequately account for the phenomenon of worker retirement. It would be desirable to expand the model to produce separate estimates of the career pattern of employment and the career path of earnings, conditional on employment. Some workers leave the labor force at a comparatively young age as a result of disability or early retirement. These workers may have rising earnings up through the point they leave the labor force. In a single-equation model of earnings, the effects of the labor market withdrawal of these early retirees is combined with the effects of continued earnings gains among workers who remain employed. The estimates of the β_A will provide reasonable estimates of the path of *unconditional* earnings, that is, earnings of workers and nonworkers alike. Unfortunately, they will obscure the potentially distinctive path of average earnings of those workers who remain employed. Equally important, they fail to reflect the abrupt drop in earnings that often accompanies worker retirement or disability.

Although we attempted to estimate a joint model that predicts employment status and average earnings conditional on employment, we encountered two problems implementing the model for purposes of making predictions of future earnings. First, the estimates of the employment equation did not produce very reliable predictions of employment. Unless we used information about each person's actual employment status in the past one or two years, we did not reliably predict the person's employment status in subsequent periods. While it might seem logical to modify the basic employment specification to include additional information about each person's actual employment status in past periods, we do not think this modification would be appropriate without thorough specification tests. Unless we can be confident that we know the correct specification of the effect of past employment status on current status, it is dangerous to make long-range predictions of future employment status based on a specification that includes lagged employment status. (This is true whether the specification explicitly includes past employment status as a regressor or it includes an auto-regressive specification of the disturbance term.) Including such lagged employment information in the specification is helpful in producing reasonably accurate -- though possibly biased -- predictions of employment status in the next period, or even in the next three or four periods. But small misspecification errors can generate large and systematic prediction errors in longer term forecasts. (In this project, we make predictions 25 or more years into the future for some of the youngest sample members.) To minimize the possibility of large out-of-sample prediction errors, analysts should closely investigate the proper time-series specification of the employment equation. Given the time and resource limits of this project, we did not think this was feasible.

A second forecasting problem associated with the two-equation approach to estimation arises because of the logical relationship between the employment-prediction and earnings-prediction equations. The estimated employment-prediction equation explains less than 100 percent of the actual variation in employment status. From the estimated employment equation we can generate predictions of future employment status over the next one to twenty-five years by using a sequence of random numbers to determine whether an individual has covered earnings in successive future years. This prediction method often produces the prediction that a person who has a very low probability of employment -- and very low or negative expected earnings -- will nonetheless be employed. The problem of producing a reasonable prediction of earnings for such an individual is formidable unless the employment-prediction and earnings-prediction equations have been simultaneously estimated, an undertaking that is well beyond the scope of this project.

3. Estimation Procedures

The earnings equation is estimated with data from the 1990-1993 Survey of Income and Program Participation (SIPP) panels matched to Social Security Summary Earnings Records (SER). The sample consisted of 44,792 women and 40,794 men for whom matched SIPP and SER records could be obtained. The sample was restricted to SIPP respondents in the 1990-1993 waves who completed the second periodic interview. (By implication the sample of "full responders" to the SIPP interviews – persons who completed all interviews that were offered to them – represents a subsample of the respondents to the second periodic interview.) The sample was further restricted to persons born between 1926 and 1965.[1]

The SER records contain information on Social-Security-covered earnings by calendar year for the period from 1951 through 1996. These records do not contain information about *all* labor earnings, but only on earnings up to the taxable wage ceiling. Censoring at the taxable maximum wage is a major problem for men in the sample, though not for women. According to our tabulations of the estimation sample, less than 1 percent of the person-year observations of women in the sample are affected by censoring. (For example, women attained the taxable maximum earnings less than 1 percent of the time between 1974 and 1983.) The problem is much more serious for men in the sample. Men's Social Security covered earnings were affected by censoring in about 15 percent of person years between 1974 and 1983. Among men born between 1921 and 1960 who were at least 22 years old, 23 percent earned wages above the taxable maximum at least once between 1984 and 1993 (when the taxable maximum was higher) and 13 percent earned wages above the taxable maximum at least once between 1994 and 1996. Men with above-average expected earnings -- for example, college graduates between 35 and 55 years old -- face a high likelihood of reaching the taxable maximum in a given year.

Censoring would not be a concern if the taxable maximum remained relatively constant. Unfortunately, it increased over the analysis period, possibly giving rise to an upward bias in estimates of the growth rate in earnings for men who have high expected earned incomes. Although we did not implement a formal censoring model, we thought it would be useful to take account of the censoring problem in a less formal and less costly way (though only in the case of males). As part of the work on stylized earnings profiles reported in Chapter 8, we created estimates of "expected earnings above the taxable maximum, but below 2.46 times average economy-wide earnings" for all men with Social Security covered earnings at the taxable maximum. For brevity, we shall refer to this transformed measure of earnings as "less censored" earnings. (This measure of earnings is also used in Chapter 8 of the project, where it was originally developed for analysis of stylized lifetime earnings patterns.)

In adjusting the censored earnings data, we did not alter the wage data for years after 1989, nor did we alter any wage reports when the reported wage was below the taxable ceiling. Starting in 1990, the Social Security taxable maximum reached 2.46 times average earnings, where it has remained. We adjusted the pre-1990 wage reports to reflect a hypothetical wage ceiling equivalent to the average wage ceiling of the 1990-96 period -- that is, a ceiling equal to 2.46 times average earnings.

For earnings in the 1951-77 period, the SER contains information on the quarter in which an individual's wages reached the taxable ceiling. This information is used to impute annual earnings for men at the taxable wage ceiling under the following rules:

Quarter reached maximum	Range of potential earnings (multiples of taxable maximum)	Predicted mean of class
4	$1 < w < 4/3$	1.14
3	$4/3 < w < 2$	1.53
2	$2 < w < 4$	2.36
1	$4 < w$	5.00

The first column shows the calendar quarter in which an individual is known to have attained the taxable wage ceiling. The second shows the probable earnings range of the individual under the assumption that he earns steady wages throughout the year. For example, a worker who attains the taxable maximum in the fourth quarter might have attained the maximum on the last day of the quarter (in which case he earned exactly the ceiling wage) or on the first day of the quarter (in which case he earned 4/3 times the ceiling wage). Given this estimate of the potential earnings range of each worker, we then derived an estimate of his expected earnings if his earnings were in the predicted range. The class means were derived from the observed distribution of wages in the Current Population Surveys (CPS) of 1965, 1970, 1975. The estimated class means were very similar for all three survey years. These average values were used to impute wages to workers above the taxable maximum for all of the years between 1951 and 1977. The resulting wage values were truncated at a value of 2.46 times the economy-wide average wage to make them consistent in their expected value with the reported data for 1990-96.

For the period 1978-89, the CPS of each year was used to obtain information on the distribution of wages in excess of that year's taxable maximum. Those wage distributions were truncated at 2.46 times the average wage, and the resulting expected values used to compute an average wage in excess of each year's taxable maximum but below 2.46 times average earnings. That conditional average wage was used in place of the value of the ceiling wage.

Once we obtained these estimates of earnings for men at the taxable wage ceiling, we still had to decide how they should be used in estimation and prediction. We chose to include "less censored earnings" as the dependent variable in an earnings regression otherwise specified in the same way as our standard earnings regression. We then compared the predictive power of the resulting estimates with those of the standard regression equation (i.e, the equation estimated on Social Security covered earnings censored at the taxable wage ceiling). The average absolute prediction error is somewhat smaller using results obtained using "less censored earnings."[2]

III. ESTIMATES AND EARNINGS FORECASTS

The dependent variable in the estimation is the worker's annual Social-Security-covered earnings divided by the economy-wide average wage for the relevant year. This ratio, which is designated y_i in equation 1, is multiplied by 100 to convert it into percentage terms. For men in the sample, "less censored earnings" is substituted for Social-Security-covered earnings in calculating the earnings ratio.

The period used in estimation is 1987 through 1996, the last ten years of available earnings data on the SER. Since the SER records cover wages earned back through 1951, we experimented with longer estimation periods. However, we have little confidence in the predictions generated using a substantially longer estimation period. Between 1973 and the present, American workers have experienced dramatic changes in lifetime earnings patterns. The gap between low-, middle-, and high-wage workers increased significantly after 1979. Pay differentials between women and men narrowed sharply. Wages of young workers fell noticeably in comparison with wages paid to middle-aged and older workers. These trends have slowed or leveled off since the late 1980s. When the estimation period includes the ten years from 1977-1986 as well as later years, the estimated coefficients imply that many of the trends observed in the late 1970s and early 1980s will continue into the indefinite future. We do not think this prediction is plausible. For that reason, we restricted the estimation period to the years since 1986, when many earnings patterns have stabilized. For each birth cohort included in the sample, the 10-year estimation period allows each cohort to move between at least two and possibly as many as six age categories defined in the age-earnings function, f(Age).

1. Coefficient Estimates

The basic earnings equation was separately estimated in eight different samples, defined by gender and educational attainment. Respondents in the highest two educational attainment groups were combined into a single estimation sample; the other three educational groups were included in separate estimation samples. The coefficient estimates, their standard errors, and 95-percent confidence intervals are displayed in Tables 2-1 and 2-2, which contain results for women and men, respectively.

Since separate age-earnings profiles are estimated for college graduates and people with post-college education, we estimate a total of 10 earnings profiles, five for women and five for men. The average estimated age-earnings profiles are displayed in Figure 2-3. The top panel shows the age-earnings profiles for five educational classes of women; the lower panel shows the profiles for men. Note that men and women with greater educational attainment have significantly higher earnings than lower education groups at all ages past about age 30. Their peak career earnings are also achieved somewhat later in life.[3] These estimates imply that relative earnings begin to decline for men between ages 40 and 50. Among men with the least

Table 2-1
Female Age-Earnings Profiles, by Educational Attainment
Fixed-Effect Model Estimates

```
Education = Less than four years of high school.
                                             Fixed-effects (within) regression
sd(u_id)                     =  31.40634              Number of obs =   74357
sd(e_id_t)                   =  17.63622                          n =    7687
sd(e_id_t + u_id)            =  36.01936                      T-bar = 9.67308
corr(u_id, Xb)               =    0.0115              R-sq within   =  0.0295
                                                           between  =  0.0309
                                                           overall  =  0.0277
                                                      F( 13, 66657) =  155.80
                                                           Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -12.29271   .7160332    -17.168   0.000       -13.69613   -10.88928
   Age29 |  -7.413093   .4267999    -17.369   0.000        -8.24962   -6.576565
   Age34 |  -3.989375   .3190682    -12.503   0.000       -4.614749   -3.364002
   Age44 |   1.516319   .3318492      4.569   0.000        .8658951    2.166744
   Age49 |   1.498146   .4367033      3.431   0.001        .6422076    2.354084
   Age54 |  -.8577039    .526687     -1.628   0.103        -1.89001    .1746024
   Age57 |  -3.946682    .608706     -6.484   0.000       -5.139746   -2.753619
   Age59 |  -6.165832   .6629575     -9.300   0.000       -7.465228   -4.866435
   Age61 |  -8.670656   .6858993    -12.641   0.000       -10.01502   -7.326294
   Age62 |  -11.72871   .7600354    -15.432   0.000       -13.21837   -10.23904
   Age64 |  -16.46597   .7275347    -22.633   0.000       -17.89194      -15.04
   Age65 |  -19.38252   .8203293    -23.628   0.000       -20.99036   -17.77467
   Age67 |  -21.67292   .8536696    -25.388   0.000       -23.34611   -19.99973
   _cons |   27.08451    .323921     83.615   0.000        26.44963     27.7194
------------------------------------------------------------------------------
      id |       F(7686,66657) =     31.175
```

```
Education = Four years of high school.
sd(u_id)                     =  43.81776              Number of obs =  174680
sd(e_id_t)                   =  22.76773                          n =   17769
sd(e_id_t + u_id)            =  49.37981                      T-bar =  9.8306
corr(u_id, Xb)               =    0.0397              R-sq within   =  0.0279
                                                           between  =  0.0353
                                                           overall  =  0.0292
                                                      F( 13,156898) =  346.19
                                                           Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -10.89394   .5343841    -20.386   0.000       -11.94132   -9.846557
   Age29 |  -7.162824   .3191782    -22.441   0.000       -7.788407   -6.537241
   Age34 |  -4.314175   .2373817    -18.174   0.000       -4.779438   -3.848912
   Age44 |   3.997877   .2495359     16.021   0.000        3.508791    4.486962
   Age49 |   5.875282   .3386881     17.347   0.000        5.211461    6.539104
   Age54 |   4.455451    .424028     10.507   0.000        3.624365    5.286537
   Age57 |   1.000265   .5142488      1.945   0.052       -.0076523    2.008181
   Age59 |  -2.859324   .5814337     -4.918   0.000       -3.998922   -1.719726
   Age61 |  -6.803079    .615085    -11.060   0.000       -8.008633   -5.597525
   Age62 |  -12.33316   .7109171    -17.348   0.000       -13.72654   -10.93978
   Age64 |  -20.02395   .6727037    -29.766   0.000       -21.34243   -18.70546
   Age65 |  -24.95934   .8005149    -31.179   0.000       -26.52834   -23.39035
   Age67 |  -27.52117   .8508359    -32.346   0.000        -29.1888   -25.85355
   _cons |   46.38095   .2162292    214.499   0.000        45.95714    46.80475
------------------------------------------------------------------------------
      id |     F(17768,156898) =     36.405
```

Table 2-1 (continued)

```
Education = One to three years of college.
sd(u_id)                     =  52.99556                Number of obs =    95846
sd(e_id_t)                   =  28.36678                            n =     9687
sd(e_id_t + u_id)            =  60.10993                        T-bar = 9.89429
corr(u_id, Xb)               =   -0.0288                R-sq within   =  0.0211
                                                             between  =  0.0113
                                                             overall  =  0.0114
                                                        F( 13, 86146) =  143.02
                                                             Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -16.69987   .8274249    -20.183   0.000      -18.32162   -15.07813
   Age29 |  -8.444023   .4972576    -16.981   0.000      -9.418643   -7.469402
   Age34 |  -4.359475   .3686321    -11.826   0.000      -5.081991   -3.636959
   Age44 |    7.01602   .3840808     18.267   0.000       6.263225    7.768815
   Age49 |   10.95157   .5347588     20.479   0.000       9.903446    11.99969
   Age54 |   11.76471   .7147145     16.461   0.000       10.36387    13.16554
   Age57 |   9.532564   .9227685     10.330   0.000       7.723946    11.34118
   Age59 |   4.897057    1.08578      4.510   0.000       2.768937    7.025177
   Age61 |  -.1251351    1.16581     -0.107   0.915      -2.410112    2.159842
   Age62 |  -6.855312   1.392399     -4.923   0.000      -9.584402   -4.126221
   Age64 |  -12.16573   1.306853     -9.309   0.000      -14.72715   -9.604307
   Age65 |  -18.95678   1.614956    -11.738   0.000      -22.12208   -15.79148
   Age67 |  -23.27936   1.722018    -13.519   0.000       -26.6545   -19.90422
   _cons |   59.27902   .3021222    196.209   0.000       58.68687    59.87118
------------------------------------------------------------------------------
      id |       F(9686,86146) =     33.950
```

```
Education = Four or more years of college.
sd(u_id)                     =  71.29666                Number of obs =    95633
sd(e_id_t)                   =  36.67594                            n =     9649
sd(e_id_t + u_id)            =  80.17691                        T-bar = 9.91118
corr(u_id, Xb)               =   -0.0931                R-sq within   =  0.0412
                                                             between  =  0.0018
                                                             overall  =  0.0058
                                                        F( 26, 85958) =  141.94
                                                             Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -36.01407   1.235241    -29.156   0.000      -38.43513   -33.59301
   Age29 |  -6.185295   .7504073     -8.243   0.000      -7.656087   -4.714503
   Age34 |  -3.943033   .5617955     -7.019   0.000      -5.044147   -2.841919
   Age44 |   6.572365   .5919927     11.102   0.000       5.412064    7.732666
   Age49 |   13.72039   .8280236     16.570   0.000       12.09747     15.3433
   Age54 |   16.02367   1.136902     14.094   0.000       13.79535    18.25199
   Age57 |   14.26148   1.495558      9.536   0.000        11.3302    17.19276
   Age59 |   9.550628   1.769274      5.398   0.000       6.082866    13.01839
   Age61 |   2.559956   1.899944      1.347   0.178      -1.163918     6.28383
   Age62 |  -5.300336   2.323021     -2.282   0.023      -9.853437   -.7472344
   Age64 |  -15.21764   2.173822     -7.000   0.000      -19.47832   -10.95697
   Age65 |  -23.69783    2.70219     -8.770   0.000       -28.9941   -18.40156
   Age67 |  -32.56209   2.897974    -11.236   0.000      -38.24209   -26.88208
Ag24_Ed5 |  -38.57714    2.69157    -14.333   0.000      -43.85259   -33.30169
Ag29_Ed5 |  -21.28159   1.529356    -13.915   0.000      -24.27911   -18.28406
Ag34_Ed5 |   -5.72276   1.086691     -5.266   0.000      -7.852666   -3.592854
Ag44_Ed5 |   2.601815   1.018081      2.556   0.011       .6063849    4.597244
Ag49_Ed5 |   2.579371    1.39273      1.852   0.064      -.1503689     5.30911
Ag54_Ed5 |   5.545336   1.871429      2.963   0.003       1.877352     9.21332
Ag57_Ed5 |   3.822669   2.504238      1.526   0.127      -1.085616    8.730954
Ag59_Ed5 |  -3.725834   2.981212     -1.250   0.211      -9.568984    2.117316
Ag61_Ed5 |  -4.334154   3.249751     -1.334   0.182      -10.70364     2.03533
Ag62_Ed5 |  -5.003977   3.939414     -1.270   0.204       -12.7252    2.717242
Ag64_Ed5 |  -12.50108    3.70212     -3.377   0.001      -19.75721    -5.24496
Ag65_Ed5 |  -14.97261   4.604314     -3.252   0.001      -23.99703   -5.948195
Ag67_Ed5 |  -10.07044   4.890624     -2.059   0.039      -19.65602   -.4848578
   _cons |   82.99281   .3778786    219.628   0.000       82.25217    83.73345
------------------------------------------------------------------------------
      id |       F(9648,85958) =     36.224
```

Table 2-2
Male Age-Earnings Profiles, by Educational Attainment
Fixed-Effect Model Estimates

```
Education = Less than four years of high school.
                                             Fixed-effects (within) regression
sd(u_id)                     =   56.7756              Number of obs =    68975
sd(e_id_t)                   =  31.80853                          n =     7140
sd(e_id_t + u_id)            =  65.07881                      T-bar = 9.66036
corr(u_id, Xb)               =   -0.2280              R-sq within   =  0.1053
                                                           between  =  0.0235
                                                           overall  =  0.0304
                                                      F( 13, 61822) =  559.40
                                                           Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -18.09379   1.254757    -14.420   0.000      -20.55312   -15.63446
   Age29 |  -7.052151   .7727004     -9.127   0.000      -8.566645   -5.537656
   Age34 |  -2.235293   .5886969     -3.797   0.000       -3.38914   -1.081446
   Age44 |  -.7398771   .6387606     -1.158   0.247      -1.991849    .5120951
   Age49 |  -5.870017   .8451327     -6.946   0.000      -7.526479   -4.213555
   Age54 |  -13.09429   1.012529    -12.932   0.000      -15.07885   -11.10973
   Age57 |  -26.52734   1.164159    -22.787   0.000      -28.80909   -24.24558
   Age59 |  -34.78413   1.263135    -27.538   0.000      -37.25988   -32.30838
   Age61 |  -44.95629   1.305595    -34.434   0.000      -47.51526   -42.39732
   Age62 |  -56.97512   1.443249    -39.477   0.000      -59.80389   -54.14635
   Age64 |  -76.24629   1.382756    -55.141   0.000      -78.95649   -73.53608
   Age65 |   -88.4447   1.571482    -56.281   0.000      -91.52481   -85.36459
   Age67 |  -95.51312   1.626268    -58.731   0.000      -98.70061   -92.32563
   _cons |   78.85383   .6119618    128.854   0.000       77.65439    80.05328
------------------------------------------------------------------------------
      id |       F(7139,61822) =     28.748
```

```
Education = Four years of high school.
sd(u_id)                     =  64.88506              Number of obs =  140285
sd(e_id_t)                   =  35.38793                          n =   14230
sd(e_id_t + u_id)            =   73.9079                      T-bar =  9.8584
corr(u_id, Xb)               =   -0.1640              R-sq within   =  0.0756
                                                           between  =  0.0291
                                                           overall  =  0.0308
                                                      F( 13,126042) =  792.46
                                                           Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -19.81902   .8722068    -22.723   0.000      -21.52853   -18.10951
   Age29 |  -7.842719   .5171317    -15.166   0.000      -8.856288    -6.82915
   Age34 |  -.9857611   .3810928     -2.587   0.010      -1.732696   -.2388258
   Age44 |  -1.930576   .4260428     -4.531   0.000      -2.765613    -1.09554
   Age49 |  -7.554708     .60333    -12.522   0.000      -8.737225   -6.372192
   Age54 |  -17.07835   .7593425    -22.491   0.000      -18.56665   -15.59005
   Age57 |  -30.66509   .9234691    -33.206   0.000      -32.47507   -28.85511
   Age59 |  -44.29774   1.055965    -41.950   0.000      -46.36741   -42.22806
   Age61 |  -59.72219   1.122345    -53.212   0.000      -61.92197   -57.52242
   Age62 |  -75.31036   1.314187    -57.306   0.000      -77.88614   -72.73457
   Age64 |  -94.51296   1.242308    -76.079   0.000      -96.94786   -92.07805
   Age65 |  -109.1274    1.49186    -73.149   0.000      -112.0514   -106.2034
   Age67 |  -117.2749   1.579743    -74.237   0.000      -120.3712   -114.1786
   _cons |   107.1683   .3496285    306.521   0.000       106.4831    107.8536
------------------------------------------------------------------------------
      id |     F(14229,126042) =     31.504
```

Table 2-2 (continued)

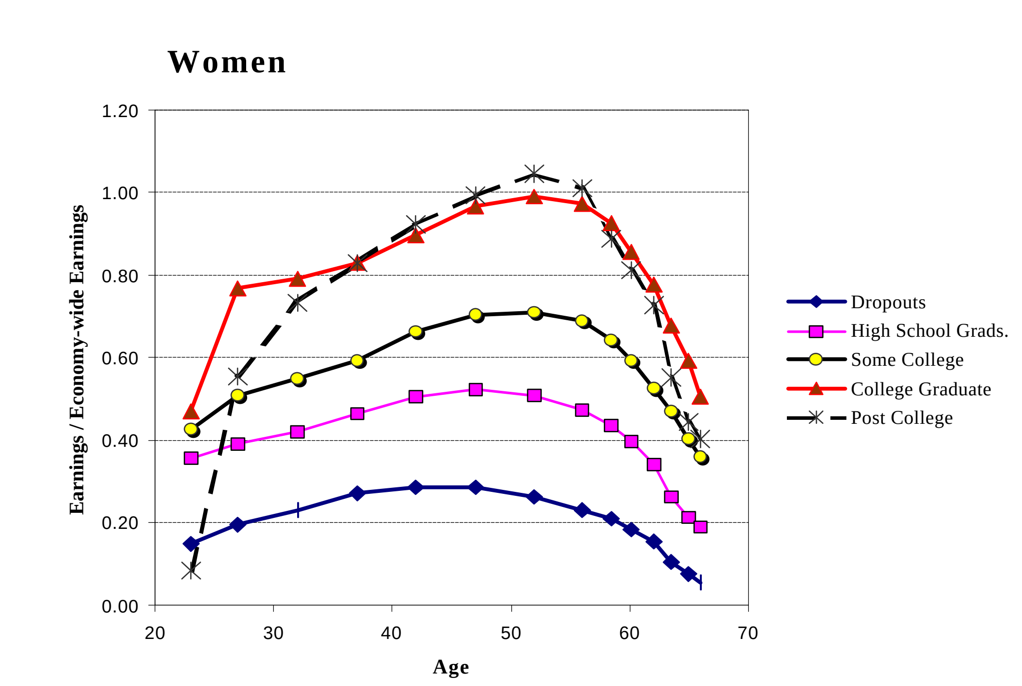
Males: Education = One to three years of college.

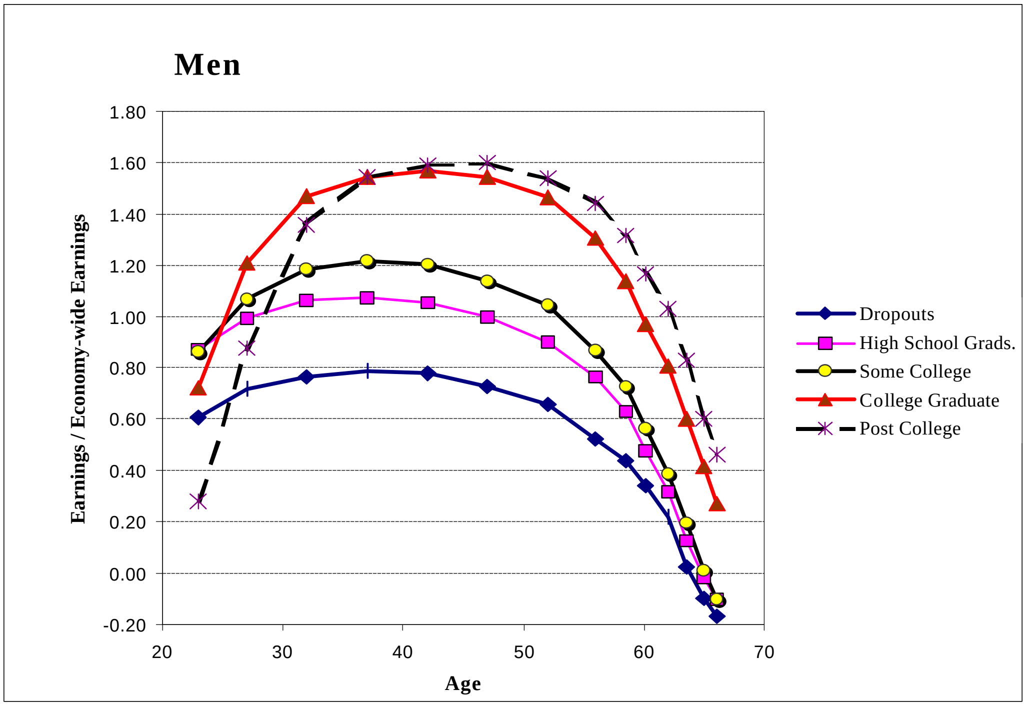
```
sd(u_id)                     =  71.31912               Number of obs =   82523
sd(e_id_t)                   =  39.21926                           n =    8332
sd(e_id_t + u_id)            =  81.39145                       T-bar = 9.90434
corr(u_id, Xb)               =   -0.1384               R-sq within   =  0.0677
                                                            between  =  0.0220
                                                            overall  =  0.0263
                                                       F( 13, 74178) =  414.62
                                                            Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -35.40504   1.269636    -27.886   0.000      -37.89352   -32.91656
   Age29 |  -14.86891   .7660261    -19.410   0.000      -16.37032    -13.3675
   Age34 |  -3.431996   .5596867     -6.132   0.000       -4.52898   -2.335012
   Age44 |  -1.634919   .5545853     -2.948   0.003      -2.721904   -.5479341
   Age49 |  -7.925679   .7744412    -10.234   0.000       -9.44358   -6.407777
   Age54 |  -17.46588   1.043255    -16.742   0.000      -19.51066   -15.42111
   Age57 |  -34.90784   1.353062    -25.799   0.000      -37.55984   -32.25584
   Age59 |  -48.89085   1.583633    -30.873   0.000      -51.99476   -45.78693
   Age61 |  -65.27757   1.696596    -38.476   0.000      -68.60289   -61.95225
   Age62 |  -82.87883   2.035584    -40.715   0.000      -86.86856   -78.88909
   Age64 |  -102.2858   1.909873    -53.556   0.000      -106.0291   -98.54241
   Age65 |  -120.5922   2.404062    -50.162   0.000      -125.3042   -115.8803
   Age67 |  -131.7692    2.57832    -51.107   0.000      -136.8227   -126.7157
   _cons |   121.7559   .4490658    271.132   0.000       120.8758    122.6361
------------------------------------------------------------------------------
      id |       F(8331,74178) =     31.343
```

Males: Education = Four or more years of college.

```
sd(u_id)                     =  80.13025               Number of obs =  109631
sd(e_id_t)                   =  44.22211                           n =   11092
sd(e_id_t + u_id)            =  91.52296                       T-bar = 9.88379
corr(u_id, Xb)               =   -0.0321               R-sq within   =  0.1027
                                                            between  =  0.0505
                                                            overall  =  0.0594
                                                       F( 26, 98513) =  433.51
                                                            Prob > F =  0.0000
------------------------------------------------------------------------------
  yratio |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Age24 |  -82.24171   1.525664    -53.906   0.000      -85.23199   -79.25143
   Age29 |  -33.71748   .9484911    -35.549   0.000      -35.57651   -31.85844
   Age34 |  -7.874092   .7164915    -10.990   0.000      -9.278407   -6.469777
   Age44 |   2.157216   .7038325      3.065   0.002       .7777123    3.536719
   Age49 |  -.0161685   .9500128     -0.017   0.986      -1.878182    1.845845
   Age54 |   -8.21484   1.280029     -6.418   0.000      -10.72368   -5.705999
   Age57 |  -24.02843   1.673674    -14.357   0.000      -27.30881   -20.74805
   Age59 |  -40.89005   1.952824    -20.939   0.000      -44.71756   -37.06254
   Age61 |  -57.72099   2.093298    -27.574   0.000      -61.82383   -53.61815
   Age62 |  -74.04065   2.519158    -29.391   0.000      -78.97817   -69.10313
   Age64 |  -94.50082   2.361653    -40.015   0.000      -99.12963   -89.87201
   Age65 |  -113.0157   2.929603    -38.577   0.000      -118.7577   -107.2737
   Age67 |  -127.6937   3.144102    -40.614   0.000      -133.8561   -121.5313
Ag24_Ed5 |  -44.49208   3.286473    -13.538   0.000      -50.93353   -38.05063
Ag29_Ed5 |  -33.26033   1.767411    -18.819   0.000      -36.72443   -29.79623
Ag34_Ed5 |   -10.8658   1.247064     -8.713   0.000      -13.31003   -8.421567
Ag44_Ed5 |   2.400491   1.154966      2.078   0.038       .1367715     4.66421
Ag49_Ed5 |   5.193227    1.52304      3.410   0.001       2.208087    8.178367
Ag54_Ed5 |   7.662023    1.97888      3.872   0.000       3.783441     11.5406
Ag57_Ed5 |    13.7329   2.556489      5.372   0.000       8.722213    18.74359
Ag59_Ed5 |    17.8258   2.980079      5.982   0.000       11.98488    23.66672
Ag61_Ed5 |   19.80614   3.193394      6.202   0.000       13.54712    26.06515
Ag62_Ed5 |   22.54329   3.834117      5.880   0.000       15.02846    30.05811
Ag64_Ed5 |   23.14078   3.602377      6.424   0.000       16.08016    30.20139
Ag65_Ed5 |   18.61051   4.451451      4.181   0.000       9.885717     27.3353
Ag67_Ed5 |   19.18038   4.770205      4.021   0.000       9.830832    28.52992
   _cons |   154.5912   .4682512    330.146   0.000       153.6735     155.509
------------------------------------------------------------------------------
      id |      F(11091,98513) =     32.151
```

Figure 2-3
Estimated Age-Earnings Profiles
By Sex and Educational Attainment





schooling attainment, relative earnings begin to fall as early as age 40. Men who have completed college do not experience sizable relative earnings declines until their 50s. Earnings peak at a lower level but at a later age among women. Peak lifetime earnings are only slightly higher than the economy-wide average wage for women with college and post-graduate educations. In contrast, among men with similar educational levels, peak earnings are approximately 60 percent higher than economy-wide earnings. Whereas men experience sizable or at least modest drops in average earnings by age 55, well-educated women do not attain their peak lifetime earnings until their middle 50s. Bear in mind that the age-earnings profiles displayed in Figure 2-3 show the combined effects of changing annual earnings among people who continue to work full time as well as steep earnings reductions associated with disability and early retirement for workers affected by these phenomena. If the estimates were based solely on earnings patterns among men and women who continue to work full time, we would see a later and higher peak in lifetime earnings.

2. Adjustments for Disability Onset

RAND Corporation analysts associated with this project generated two kinds of predictions that affect our projections of future earnings. They produced both a prediction of the onset of a health problem that limits the kind or amount of work a person can do and a prediction of the calendar year of death. The latter prediction was used to zero out predicted Social Security covered earnings for all years after the predicted date of death. RAND’s prediction of a health limitation was used to help predict the onset of Social Security Disability Insurance (DI) receipt. In this section, we explain how our estimates of DI onset were obtained and how they are used to modify our forecast of earnings for people predicted to receive DI pensions.

We used data in the Social Security Administration’s Master Beneficiary Record (MBR) to derive an estimate of the onset of DI payments for matched SIPP-SER sample members. These estimates range back to 1957 (when the DI program was established) up through 1998. Because the MBR data show an unexpected decline in the incidence of new DI awards beginning in 1995, the MBR file does not fully reflect new DI awards in the 1995-98 period. We therefore used the data for the calendar years 1987 - 1994 as a basis for estimating a Probit equation that predicts DI onset.

The Probit coefficients are displayed in Table 2-3, titled “Probit Model of Disability Insurance Onset by Gender, 1987 - 1994.” The age category variables are the same as those described above. In addition, the independent variables include race or ethnicity indicator variables (“black” and “whhis” -- white Hispanic), an indicator variable (“disabl”) derived from RAND’s prediction of the onset of a health problem that limits the kind or amount of work a person can do (set equal to zero in years before RAND predicts a health limit and set equal to one in later years), educational attainment indicator variables (“Edc1” through “Edc5”) associated with five levels of schooling (less than four years of high school, one to three years of college, four years of college, and five or more years of college; the *omitted* category is four years of high school), and indicator variables ("avernc1" through "avernc6" or "avernc8") that reflect the person's average Social Security covered earnings in the 10-year period ending in the calendar year before the year in the estimation.

We developed our specification of the effect of past indexed earnings after some experimentation with alternative approaches. Our first approach was to attempt to measure precisely the eligibility status (except for level of health impairment) of each person in the sample. According to the Social Security Act, a person's eligibility is determined under a two-part test that involves the person's total credited quarters of covered earnings and the level of covered earnings in the recent past. To be eligible for a DI pension, a person suffering serious health impairment must meet both these tests. We tried to apply the tests for each year in the estimation period based on earnings information in the SER. According to our calculations, there were a handful of people who began to receive DI pensions even though they did not pass both tests. It is of course possible that our program did not accurately reflect the two-part test for eligibility. It is probable that the person's eligibility was determined on the basis of earnings received in a period different from the one we assumed.

The failure of our program to distinguish accurately between eligible and ineligible workers led us to take a different approach to the specification of DI onset. Workers with no or very low covered earnings in the recent past should be ineligible for DI benefits. However, eligible workers with moderately low earnings are often found to have the highest propensity to apply for benefits. There are two likely reasons for this. First, workers with low recent earnings have low potential earnings. Under the redistributive formula that determines DI pensions, these workers receive benefits that are very generous relative to their potential earnings. The high replacement rate makes it financially more attractive for low-potential-earnings workers to apply for DI. Second, a disproportionately large percentage of low-wage jobs are in manual occupations with physically demanding work requirements. Health impairments are more likely to make it impossible or very unpleasant to continue to work in these jobs. A reasonable specification of the effect of lagged past earnings is that lagged earnings will have a nonlinear effect on the probability of DI onset. Zero and very low past earnings levels should make DI onset very improbable, because the person is not likely to meet the two-part earnings requirements. Somewhat higher past earnings levels should be associated with above-average probability of DI onset. Further increases in lagged earnings above some threshold level should be associated with declining probability of DI onset.

Our final specification of DI onset reflects this reasoning. We divided 10-year-lagged average earnings into 6 to 8 categories. The first category represents a very low level of 10-year-average earnings (15 percent or less of economy-wide average earnings), while other categories reflect successively higher levels of 10-year-average earnings. For calendar years 1977 - 1996, actual earnings are used to derive our estimates of 10-year-average earnings; for calendar years 1997-2024, our predicted Social Security covered earnings are used.

Table 2-3
Probit Model of Disability Insurance Onset
by Gender, 1987 - 1994

Dependent variable is onset of Social Security Disability Insurance /
Sample each year consists of persons who have not begun receiving
DI as of December 31 of the previous calendar year.

Females: Probit model of DI onset.

```
(Sum of weights is  5.5827e+011)

Probit Estimates                                       Number of obs = 339323
                                                       chi2(21)      =2118.29
                                                       Prob > chi2   = 0.0000
Log Likelihood = -5165.8543                            Pseudo R2     = 0.2035

                            (standard errors adjusted for clustering on newid)
------------------------------------------------------------------------------
         |               Robust
      DI |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   black |   .1654386   .0355755      4.650   0.000       .0957119    .2351653
   whhis |  -.0848271   .0561367     -1.511   0.131      -.1948531    .0251989
  disabl |   1.074342    .030751     34.937   0.000       1.014071    1.134612
   Age34 |   .0606031   .0654301      0.926   0.354      -.0676374    .1888437
   Age39 |   .0499833   .0677031      0.738   0.460      -.0827124    .1826789
   Age44 |    .153239   .0658025      2.329   0.020       .0242684    .2822095
   Age49 |   .2056123   .0655228      3.138   0.002         .07719    .3340345
   Age54 |    .379617   .0637878      5.951   0.000       .2545952    .5046389
   Age57 |   .5084386   .0678361      7.495   0.000       .3754824    .6413948
   Age59 |    .517327   .0711581      7.270   0.000       .3778598    .6567943
   Age61 |   .4578733    .072932      6.278   0.000       .3149292    .6008174
   Age67 |   .1240795   .0825009      1.504   0.133      -.0376193    .2857784
    Edc1 |   .2479518   .0333937      7.425   0.000       .1825013    .3134023
    Edc3 |  -.0484136    .037074     -1.306   0.192      -.1210773    .0242502
    Edc4 |  -.1559035   .0530473     -2.939   0.003      -.2598744   -.0519327
    Edc5 |  -.2161441   .0749549     -2.884   0.004       -.363053   -.0692351
 avernc1 |  -.6339791   .0411518    -15.406   0.000      -.7146351   -.5533231
 avernc2 |  -.0300959   .0375582     -0.801   0.423      -.1037087    .0435169
 avernc4 |   .0144312   .0406474      0.355   0.723      -.0652362    .0940985
 avernc5 |   .0288912   .0517097      0.559   0.576       -.072458    .1302405
 avernc6 |  -.0621247    .062069     -1.001   0.317      -.1837777    .0595283
   _cons |  -3.179644   .0578709    -54.944   0.000      -3.293069   -3.066219
------------------------------------------------------------------------------
```

Table 2-3 (continued)

Males: Probit model of DI onset.

```
(Sum of weights is  5.2789e+011)

Probit Estimates                                        Number of obs = 305837
                                                        chi2(23)      =2788.75
                                                        Prob > chi2   = 0.0000
Log Likelihood = -6274.4237                             Pseudo R2     = 0.2251

                            (standard errors adjusted for clustering on newid)
------------------------------------------------------------------------------
         |               Robust
      DI |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   black |   .2338922   .0377195      6.201   0.000        .1599634    .3078209
   whhis |   .0065133   .0517531      0.126   0.900       -.0949209    .1079475
  disabl |   1.084451   .0276792     39.179   0.000        1.030201    1.138701
   Age34 |   .0032535   .0630123      0.052   0.959       -.1202483    .1267553
   Age39 |   .1608041   .0624904      2.573   0.010        .0383252    .2832829
   Age44 |   .1940483   .0627655      3.092   0.002        .0710303    .3170663
   Age49 |   .2852594   .0629421      4.532   0.000        .1618951    .4086236
   Age54 |   .4432983   .0615067      7.207   0.000        .3227475    .5638492
   Age57 |   .6222472   .0656202      9.483   0.000         .493634    .7508603
   Age59 |   .6418724   .0679842      9.441   0.000        .5086258    .7751191
   Age61 |   .6207006   .0716262      8.666   0.000        .4803157    .7610854
   Age67 |   .4473458   .0724628      6.173   0.000        .3053213    .5893702
    Edc1 |   .1511553   .0317329      4.763   0.000          .08896    .2133505
    Edc3 |   -.110161   .0362503     -3.039   0.002       -.1812103   -.0391117
    Edc4 |  -.2139388   .0486935     -4.394   0.000       -.3093763   -.1185013
    Edc5 |  -.1947839   .0616041     -3.162   0.002       -.3155257   -.0740422
  avernc1 |  -.6399623   .0554213    -11.547   0.000        -.748586   -.5313386
  avernc2 |   .0476302   .0468815      1.016   0.310       -.0442558    .1395161
  avernc4 |  -.1028289   .0400632     -2.567   0.010       -.1813513   -.0243066
  avernc5 |  -.0679954   .0417819     -1.627   0.104       -.1498865    .0138956
  avernc6 |  -.2105234   .0426067     -4.941   0.000       -.2940311   -.1270157
  avernc7 |  -.2200916   .0561161     -3.922   0.000       -.3300772   -.1101061
  avernc8 |  -.2441333   .0576828     -4.232   0.000       -.3571895   -.1310772
   _cons |  -3.125513   .0502278    -62.227   0.000       -3.223958   -3.027069
------------------------------------------------------------------------------
```

The exact definitions of the earnings categories, for women and men, respectively, are as follows:

	Range of 10-year-average earnings as % of economy-wide earnings	
Variable name	Women	Men
avernc1	0 - 15 percent	0 - 15 percent
avernc2	15 - 30	15 - 30
[*Left out category:*	*30 - 70*	*30 - 70]*
avernc4	70 - 100	70 - 100
avernc5	100 - 130	100 - 130
avernc6	130 percent or more	130 - 180
avernc7	------	180 - 210
avernc8	------	210 percent or more

The Probit equations were estimated using weights reflecting the SIPP 2nd interview panel weights. The weighted estimates are virtually identical to those obtained with unweighted estimation, but the predictions of future DI incidence using weighted estimation appeared slightly preferable to those generated with coefficients obtained with unweighted estimation. (In particular, the rise in the predicted incidence of DI as the population ages after 1994 was somewhat smoother using the weighted rather than unweighted coefficient estimates.)

We used information from the MBR to predict DI onset for those persons found to begin receiving DI benefits between 1995 and 1997, inclusive. The people with records in the MBR that show they began receiving DI between 1995-1997 were “predicted” to begin receiving benefits in those years. The probability that other persons in the sample would experience DI onset was adjusted (i.e., was reduced) to reflect the existence of those persons already “predicted” to begin receiving DI because the MBR showed that they actually began to receive DI between 1995-1997. In years starting in 1998, persons were predicted to begin receiving a DI pension based solely on the probabilities predicted by the coefficient estimates in Table 2-3. Assignment to DI status was generated by comparing the person’s predicted probability of DI onset to a randomly selected number in the unit interval. Persons with a random number *below* the probability threshold implied by the Probit prediction were identified as beginning to receive DI; persons with a random number *above* the threshold were identified as non-DI recipients for the year. This process was repeated for all calendar years before 2026 in which a person was predicted to be alive and was age 64 or less. (Because people age 65-66 were not eligible for DI benefits between 1987-1994, it is impossible to estimate reliably what fraction of them would begin collecting DI benefits if the Normal Retirement Age were raised to 67, as it will be in the next century.)

The MBR files show that 5.1 percent of males in the matched SIPP/SER sample and 3.6 percent of females in the matched sample began receiving DI pensions between 1957 and 1994 (the final year used to estimate the coefficients displayed in Table 3). Our predictions of DI onset imply that another 15.8 percent of males and 12.9 percent of females will begin receiving DI after

1994 and before they attain age 65 or die. Thus, a total of 20.9 percent of all men and 16.5 percent of all women in the matched sample are observed or predicted to collect DI benefits before reaching age 65. For both sexes combined, the cumulative probability of collecting DI is 18.6 percent, implying that the annual hazard of beginning a DI pension is about 0.47 percent. (That is, if the hazard of disability onset is 0.47 percent a year from age 22 through age 64, 18.6 percent of people who survive through age 64 will collect a DI pension.)

The predictions of DI onset were used to modify our predictions of Social Security covered earnings in years in and after the year of predicted DI onset. We reduced our estimate of predicted earnings to zero. Some workers, of course, will continue to have modest covered earnings, even after they start receiving a monthly DI check. Others will fully or partially recover from their disability and resume regular work. These earnings are suppressed by our procedure of zeroing out earnings after DI onset. It would take a more elaborate model of DI entry and exit than we have developed here to predict the actual pattern of earnings of DI beneficiaries after they have first become eligible for DI pensions.

3. Estimation Issues and Possible Extensions

The predictions of DI onset as well as the integration of these predictions with other components of the MINT model could certainly be improved. As noted above, in our estimation of DI onset we were unable to distinguish reliably between workers who were currently eligible and ineligible for DI benefits, assuming their disability was sufficiently serious. This aspect of the estimation could be improved if more time and effort were devoted to determining whether workers meet the precise covered earnings requirements for current DI eligibility in each calendar year.

The prediction of individual earnings could also be improved if health limitations and DI onset were explicitly taken into account in the estimation of the lifetime earnings function. Our current method of predicting earnings for people whom we predict to receive DI benefits may lead to a problem with the distributional characteristics of predicted earnings. Our one-equation earnings model is estimated using all respondents to the second SIPP interview who were successfully matched to Social Security Earnings Records. Thus, the coefficient estimates reflect earnings patterns in this broad sample, including the members of the sample who begin to receive DI benefits before 1995 and members predicted to receive DI benefits in or after 1995. By implication, our earnings predictions are strictly valid only for this entire sample.

After predicting the future DI status of sample members, however, we “zero out” earnings in years starting in the year people are predicted to begin receiving DI benefits. This procedure produces revised predictions of earnings that will cause an understatement of future earnings for the overall sample. The understatement does not affect future Social Security benefits under certain assumptions. In particular, if the original predicted earnings of DI recipients were accurate and if we have accurately predicted the sample members who will receive DI benefits, then the predicted earnings that have been zeroed out will have little if any impact on Social

Security benefits. That is because the zeroed-out earnings will be received by people with DI pensions, and their pensions should not be affected by the modest wages they earn after DI onset. This assumption is unlikely to be entirely accurate, however. It seems probable that many of the people we predict to receive DI benefits will not receive benefits. Their earnings (before and after the predicted onset of DI benefits) will determine OAI pensions rather than DI pensions.

The magnitude of the prediction problem depends, in part, on the amount of predicted earnings which are zeroed out. People who are predicted to receive DI benefits have much less earnings than average, even before their earnings are zeroed out. This is because the equation that predicts DI onset includes workers' health limitation and lagged earnings as independent variables. Persons with below-average lagged earnings have a much higher predicted probability of obtaining DI benefits than persons with average or above-average earnings. Thus, the extent of the prediction problem is likely to be modest.

A straightforward way to improve the predictions is to estimate a second earnings function, one that is estimated using only the sample that is predicted to survive to age 62 without collecting a DI pension. The coefficient estimates of such a function would produce earnings predictions that have somewhat higher average values than the predictions generated by the current version of the model. Of course, even this is not a fully consistent or satisfactory solution to the estimation problem. In principle, the best theoretical solution is to estimate a model in which workers' annual earnings patterns *and* the onset of worker disability are simultaneously estimated. The theoretical and statistical demands of estimating such a model are beyond the scope of this project, however. The primary goal of estimating DI onset was to remove from the main MINT sample those people predicted to receive DI benefits before attaining age 65. Given this modest goal, the estimated model is probably adequate.

IV. PATTERN OF FUTURE EARNINGS GROWTH

1. Methodology

It is straightforward to generate predictions of earnings outside of the estimation period. An estimate of the individual-specific fixed effect (μ_i) is added to estimates of $X_{it}\beta$ to produce an estimate of the person's expected covered earnings in year *t*. In order to generate predictions that have a similar variance to actual covered earnings, we also added a time-varying error term to the prediction. The error term was generated by forming estimates of each person's time-varying error term for each year between 1987 and 1996. We then randomly selected an error term from the ten estimated error terms.

The predictions are similar in some ways to those generated by using simpler estimation methods (e.g., see Iams and Sandell, 1997). As in previous attempts to measure future earnings patterns, the predictions are based on only ten years of past earnings rather than each person's full age-earnings history from age 22 through calendar year 1996. Like the earlier Iams-Sandell estimates, the predictions are based on a fairly standard age-earnings pattern for representative workers from a handful of populations, in our case defined by sex and educational attainment. Our predictions differ in a couple of respects from earlier ones, however. The cohort-specific effects are implicitly assumed to be part of the person-specific error term that is unique to each sample member. The method of predicting future annual earnings introduces substantially more year-to-year variability in post-1996 earnings. It also allows year-to-year earnings fluctuations to differ in a systematic way from one person to the next, based on the observed variability of each person's past earnings during the estimation period. Tabulations performed by the Social Security Administration suggest our predictions of future average indexed monthly earnings (AIME) are similar to those made earlier by Iams and Sandell.

2. Qualifications and Alternative Approaches

Selection of Error Terms

An alternative procedure to the one outlined above would be to select a random term from a normal distribution with mean zero and standard deviation equal to the estimated standard deviation of the time-varying error term. (For female high school dropouts, for example, this standard deviation was estimated to be 17.636 percent of economy-wide average earnings.) The procedure we used is preferable to selecting a purely random error, because it permits the variance of the time-varying error term to differ from one person to the next. Moreover, it does not impose the assumption that the time-varying error terms are normally distributed. (Statistical tests suggest that the time-varying error term is *not* normally distributed.) On the other hand, our procedure takes no account of the fact that the variance of the time-varying error changes as workers grow older. In particular, it is likely that the disturbance pattern for workers past age 55 is different from what it was when the same workers were in their 20s or 30s. In an extension of the present work, it is desirable to investigate this possibility and to adjust the imputed disturbances to reflect it.

Two Equation Model of Employment and Earnings

All of the predictions are obtained using a fixed-effect model of unconditional earnings. Although we attempted to estimate a version of a two-equation model that explained both the employment relationship and the conditional earnings function, the predictions of future earnings produced by this model proved unsatisfactory. Under the two-equation approach, we estimate a first-round *employment* equation and then a second-round *earnings* equation for persons who have positive earnings. The logic behind this approach is that much of the year-to-year variation in covered earnings, especially among women and older men, is produced by entry into or exit from covered employment (i.e., movements between no earnings at all and positive earnings). There may be much smaller variation in earnings among workers, conditional on the fact that the workers have at least $1 of covered earnings. As noted above, the estimates of the first-round employment equation produced unreliable out-of-sample predictions of employment and consequently very poor estimates of the year-to-year pattern of unconditional earnings.

Nonetheless, a useful extension of the current earnings model would take explicit account of lengthy periods of nonemployment, especially periods that begin with the worker's retirement or permanent disability.

A two-equation approach to estimating earnings offers potentially huge advantages in predicting future patterns of retirement and labor market withdrawal. But if the goal of the forecast is to predict average OAI benefits accurately, the most conservative approach is to obtain good estimates of a single-equation, unconditional earnings function and then to rely on that model for long-term predictions.

Several implications of our choice of approach should be emphasized, however. First, the method produces too few predictions of consistently low or zero earnings, especially among workers nearing typical ages of retirement. In policy simulations where the exact number of years with positive earnings is important (for example, in predicting the impact of increasing eligibility quarters for disability and old-age benefits) this shortcoming can represent a serious problem. Second, the method will produce too few predictions of non-standard age-earnings profiles. For example, few people who are age 40 or younger in 1996 will be predicted to have a "slumped" pattern of lifetime earnings, even though such a pattern occurs fairly often in practice. Third, the absence of an auto-regressive error pattern in the predictions means that the predictions of labor market withdrawal late in life will not mirror actual patterns.

Implications of Model's Treatment of Retirement

"Retirement" is generally interpreted to mean that people's earnings go to zero and then remain there. Although analysts have found that labor force re-entry after retirement is quite common, the popular conception of retirement (complete and permanent exit from the work force) is probably the dominant pattern for most workers. The prediction method used here will under-represent this dominant pattern.

As noted above, the use of a single-equation does not produce biased forecasts of adjusted indexed monthly earnings (AIME) and therefore is a reasonable method for SSA to use to forecast the distribution of Social Security retirement benefits. The failure of the forecasts to capture the increased prevalence of zero or very low earnings (combined with higher average earnings for those who do not retire) among workers between ages 55 and 61 does, however, reduce the quality of the forecasts of other sources of retirement income. In particular,the forecasts of both pension benefits of workers and of non-pension wealth accumulation between ages 55 and 61 are conditional on the pattern of earnings as well as on the AIME. The quality of these projections would be improved, perhaps significantly, if the earnings projections adequately reflected the diversity of outcomes for workers between ages 55 and 61.

It is important to emphasize, however, that MINT does not totally ignore the retirement decision. In Chapter 5, we present a model of the decision to accept Social Security retirement benefits for workers aged 62-66. Chapter 6 presents a model of partial retirement earnings for those workers who are projected to accept retirement benefits. The earnings forecasts derived in Chapters 5 and 6 for workers over age 62 who accept benefits supercede the tentative earnings forecasts for these workers that are derived from the methodology discussed in this chapter.

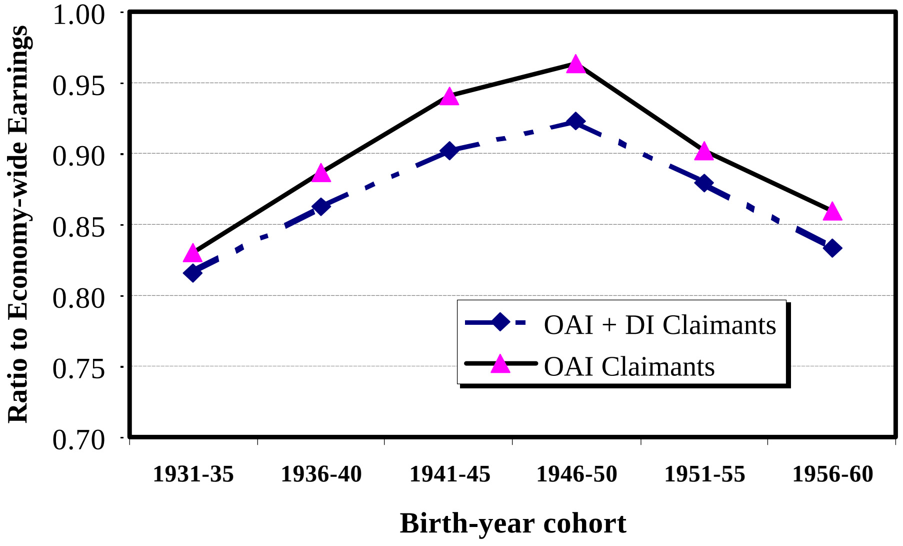
3. Projection Results

Average Lifetime Earnings

Overall, our predictions of future earnings look reasonable. Both the mean of predicted earnings and the variance of the predictions seem sensible in view of the observed trend and distribution of actual earnings over the 1974-1996 period. The calculations displayed below are based on our estimates of each worker's AIME. Our predictions of annual covered earnings are converted into indexed earnings and averaged over a career to calculate the AIME. For workers who claim an Old-Age Insurance (OAI) pension at age 62, AIME is calculated by choosing the highest 35 years of indexed earnings up through age 61 and then dividing by 35 × 12 (35 years times 12 months per year). The AIME formula for workers claiming DI pensions uses a smaller number of years in the calculation, because workers typically apply for benefits before reaching age 62, but the principle of the calculation is the same.[4] After forecasting annual earnings for 1997 and later years, we can create projected lifetime earnings histories for people in the matched SIPP-SER sample. Combining these estimates with RAND's predictions of the age of mortality and our predictions of DI onset, we can make predictions of AIME for workers who obtain enough earnings credits to become eligible for benefits.

Figure 2-4 shows the trend in predicted AIME, measured as the fraction of economy-wide earnings in the year a worker attains age 62. The tabulations cover two groups of workers. Members of both groups must have full panel weights on the 1990-1993 SIPP surveys, must

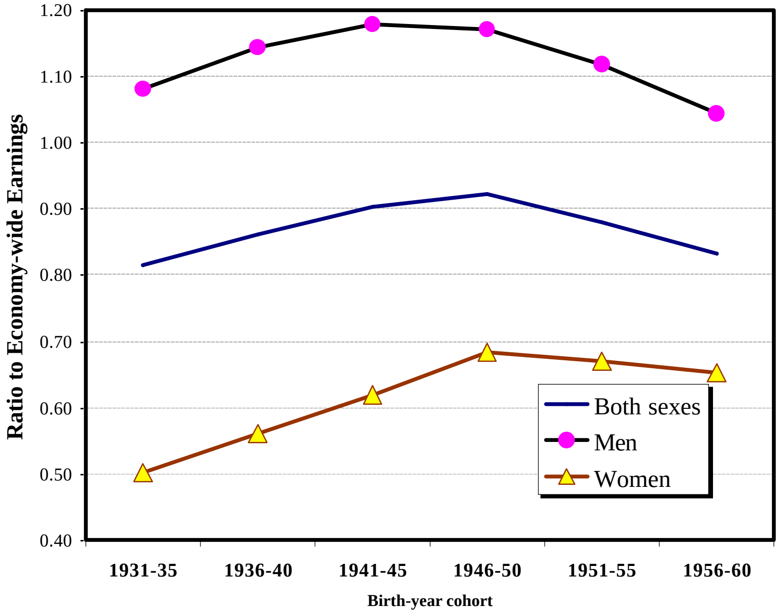
Figure 2-4
Trend in AIME by Birth-Year Cohort: Both Sexes



survive until age 62, and must accumulate enough quarters of Social Security covered earnings to become entitled to OAI or DI pensions. The lower line in the panel shows the AIME of members of the six birth cohorts who qualify for *either* a DI or an OAI pension. The upper line shows the AIME of members of this same group *except* those workers who collect a DI pension before attaining age 62. Not surprisingly, the average earnings of the latter group are higher than those of the former. Workers who receive DI pensions are disproportionately drawn from the low-wage workforce. As a result, the overall level of AIME is about 4 percent smaller in the sample that includes DI recipients than it is in the sample that includes only OAI recipients.

For both groups of workers, the average AIME increases for the first four birth cohorts and then declines for the two most recent cohorts (those born in 1951-55 and 1956-60). For the combined sample of DI and OAI claimants, the average prediction of AIME rises more than 10 percentage points of economy-wide average earnings before falling about 9 percentage points for the two most recent cohorts. Figure 2-5 shows how this overall pattern of rising and then falling average AIMEs is divided between men and women. (The calculations shown in Figure 2-5 and in all subsequent figures and tables are based on the combined sample of DI and OAI claimants.) The solid line in the middle of the figure shows the trend in overall average AIME for both sexes combined. It is identical to the dashed lower line in Figure 2-4. The top line in Figure 2-5 shows the cohort trend in AIME for men. It shows a pattern of gradually rising AIME's for the first three cohorts and somewhat sharper declines in the two most recent cohorts. The lowest line in Figure 2-5 shows the AIME trend among women. The average AIME rises sharply across the earliest four cohorts, increasing by 18 percentage points of economy-wide earnings (more than one-third) in comparison with the AIME of women born between 1931 and 1935. The average projected AIME then declines slightly for women in the two most recent cohorts.

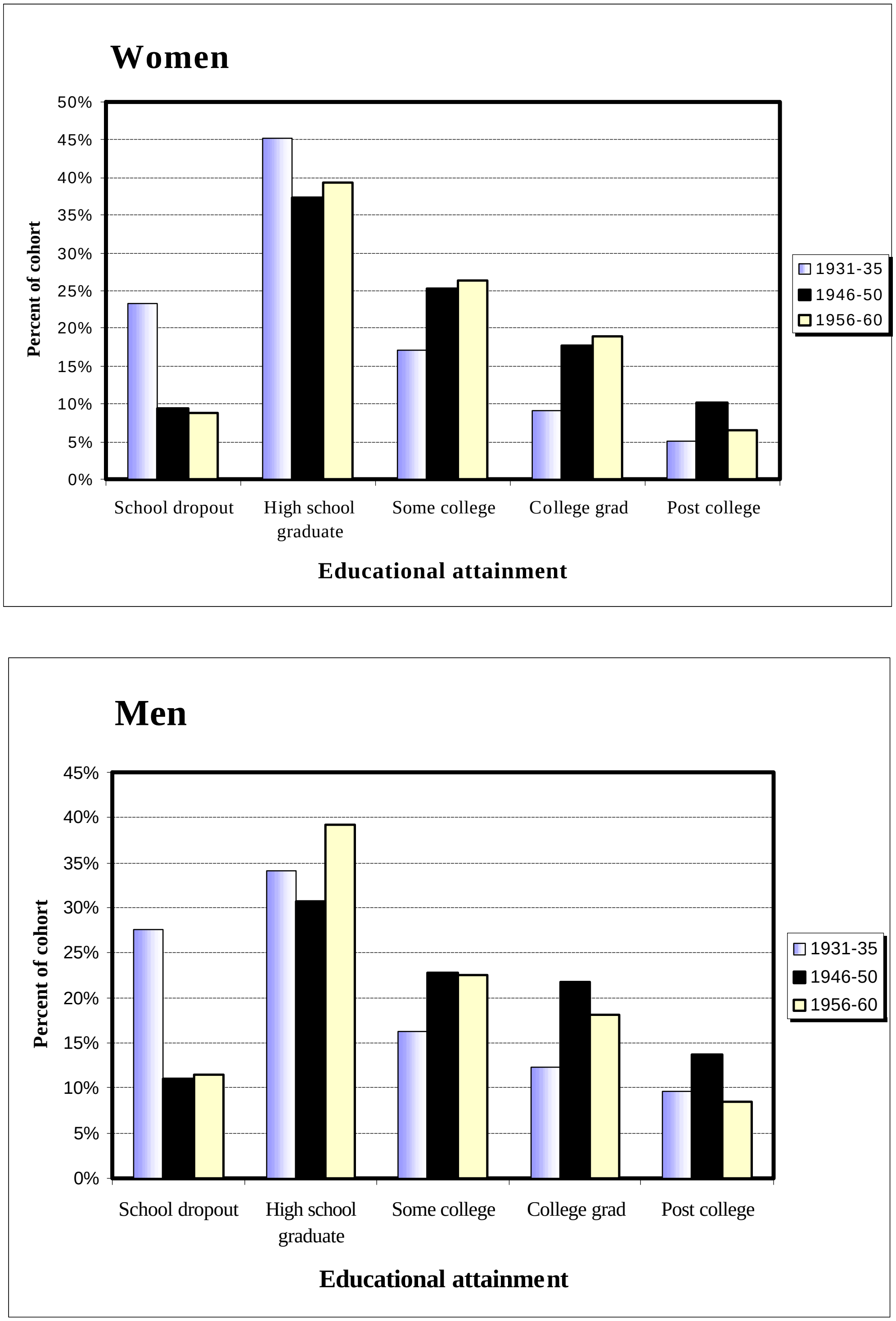
Figure 2-5
Trend in AIME by Birth-Year Cohort: Both Sexes



In part the initial rise in predicted AIMEs is explained by increasing levels of school attainment in the work force. Workers with more education enjoy a steeper gain in earnings when they are young and reach their peak earnings at later ages. Figure 2-6 shows the distribution of schooling attainment in three of the birth cohorts, the earliest (born 1931-35), the one with peak AIMEs (born 1946-1950), and the most recent (born 1956-60). The top panel shows the education distribution among women; the lower panel shows the distribution among men. There were clearly sharp drops in the proportion of workers who did not complete high school and sharp increases in the proportion with advanced levels of school attainment for the first Baby Boom cohort, born between 1946 and 1950. Gains in educational attainment are less clear for the most recent cohort. The proportions of women with some college and with a college degree or post-college education continue to increase for the most recent cohorts. For men, however, there is a drop in the proportion of workers with a post-college education and there is even a drop in the fraction who have completed college. Part of these differences may be explained by the timing of the SIPP surveys, which is the source of information on workers' educational attainment. The last SIPP interview for these workers was administered between 1992 and 1995, when workers born in 1960 were between 32 and 35 years old. In contrast, early Baby Boom workers were at least 42 to 45 years old in those years. Some schooling obtained when later Baby Boom workers were in their middle and late 30s will be missed by the SIPP interviews. Completed school attainment of the younger workers will probably be somewhat higher than these tabulations reflect. Nonetheless, it seems likely that men in the latter part of the Baby Boom generation will never attain the levels of advanced education received by early members of the Baby Boom. As a result, fewer of them may achieve the high earnings and steeper wage increases that have been received by the early Baby Boomers.

The decline in average AIMEs among more recent cohorts of workers is also the result of their low levels of relative earnings when they were young. Cross-sectional tabulations of the Current Population Survey (CPS) show that earnings of men in their 20s were sharply lower during the 1980s and early 1990s than was the case in the 1970s, especially for men with less than a college degree (see Levy and Murnane, 1992; Burtless, 1995; and Freeman, 1997). Since the *relative* earnings of these men were lower than those of earlier cohorts at the same age, these men will be predicted to have lower *relative* lifetime earnings under the assumptions of the model. The AIME is simply the unweighted average of relative earnings for the 35 years of highest relative earnings in a worker's career. If the first 10 or 15 years of a worker's career are scarred by low relative earnings, it will be impossible, under the assumptions of the model, for this poor performance to be overcome. It might be the case, of course, that workers in more recent cohorts will experience a steeper rise in their relative earnings as they move from their 20s and 30s to their 40s and 50s than earlier cohorts experienced. Alternatively, they may delay their retirement or experience smaller relative earnings declines in their 50s and 60s than was the case with earlier cohorts. But this kind of forecast seems to me no more likely than the prediction that their age-earnings profiles, after adjustment for their low *initial* earnings experiences, will be no steeper than those of earlier cohorts. By implication, the later Baby Boom workers, especially men, will earn lower relative earnings over their lifetimes than the more advantaged Baby Boom workers who were born immediately after World War II.

Figure 2-6
Distribution of Educational Attainment by Birth Cohort


Women

Men

The distribution of lifetime earnings

Our estimates can also be used to examine the distribution of lifetime earnings within cohorts. Figure 2-7 shows trends in the average AIME within fifths of the AIME distribution and for entire cohorts. The estimates suggest that average AIMEs rose within each fifth of the AIME distribution through the cohort born from 1946 to 1950 and then declined within each fifth of the AIME distribution for the two most recent cohorts. This pattern is especially pronounced at the top end of the AIME distribution, in part because the taxable maximum earnings threshold has risen sharply since the mid-1960s. The trend in AIME *inequality* may be somewhat clearer in Figure 2-8. This figure shows the average AIME in each fifth of the AIME distribution as a ratio of the average AIME in the middle fifth of the AIME distribution. (This ratio is always exactly 1.00 for the middle fifth of the distribution.) Women have a less equal distribution of AIMEs than men. The average AIME in the top fifth of the female distribution is roughly 2.5 times the average AIME in the middle of the female distribution. Among men the same ratio is roughly 1.75. At the other end of the AIME distribution, women in the bottom fifth earn about 28 percent of the average amount earned by women in the middle. Men at the bottom earn about 35 percent of the amount earned by men in the middle.

Figure 2-7
Trend in AIME by Birth Cohort and Fifths of AIME Distribution: Both Sexes

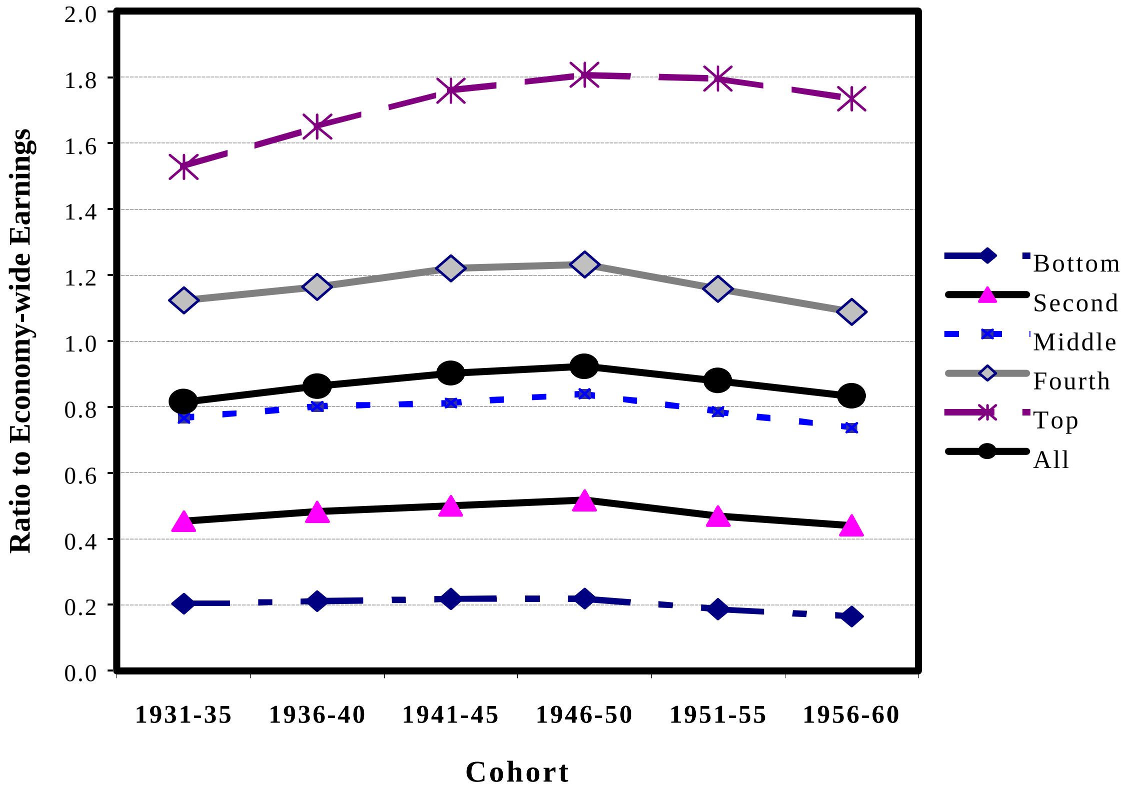
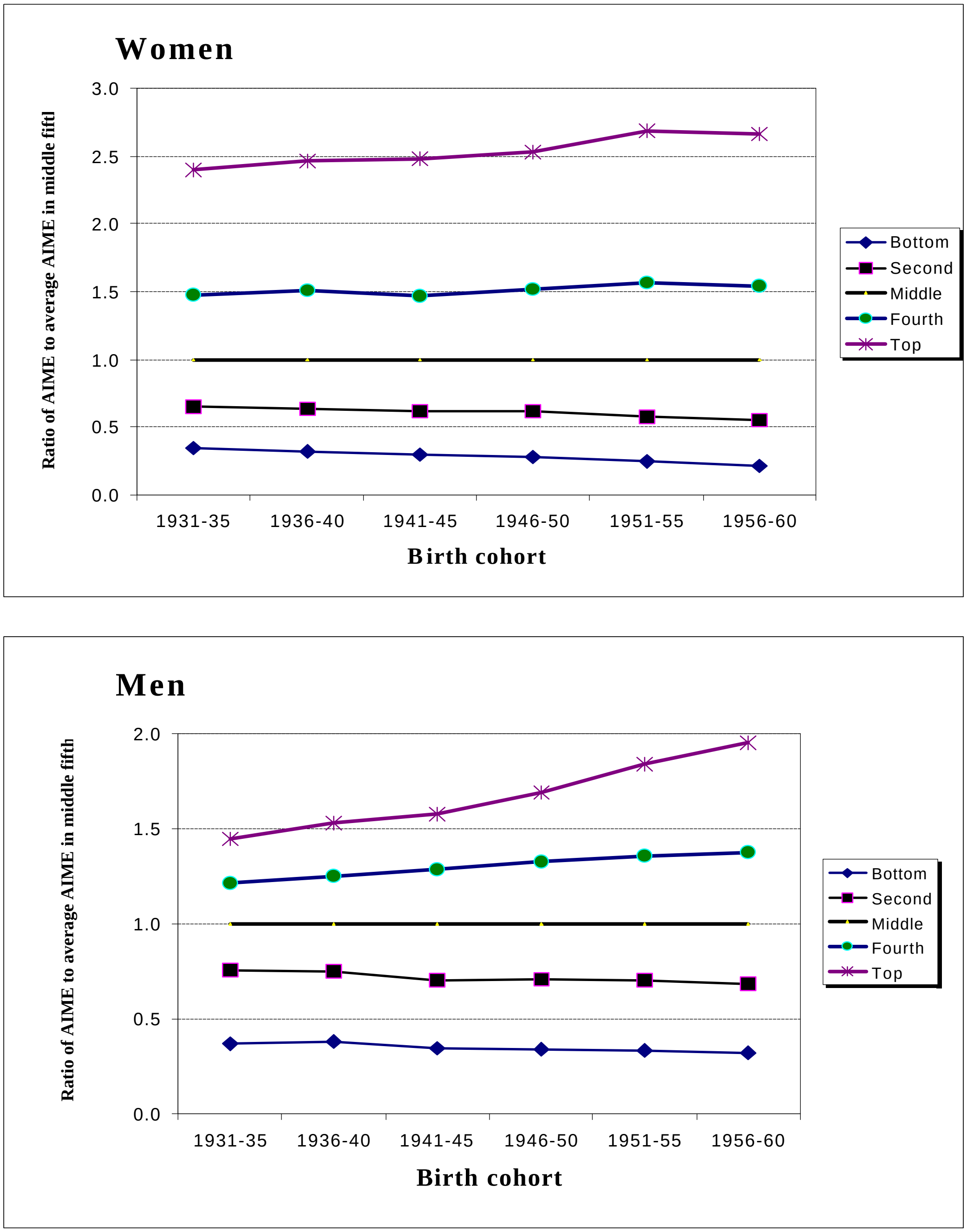


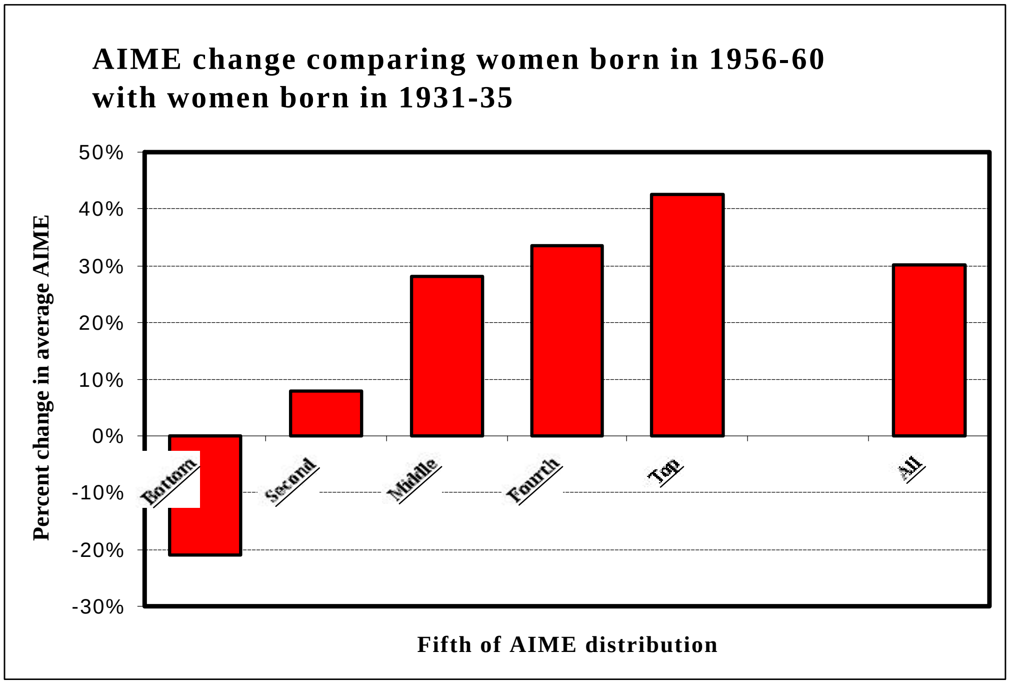
Figure 2-8
Trend in Relative AIME by Birth Cohort Within Fifths of the AIME Distribution

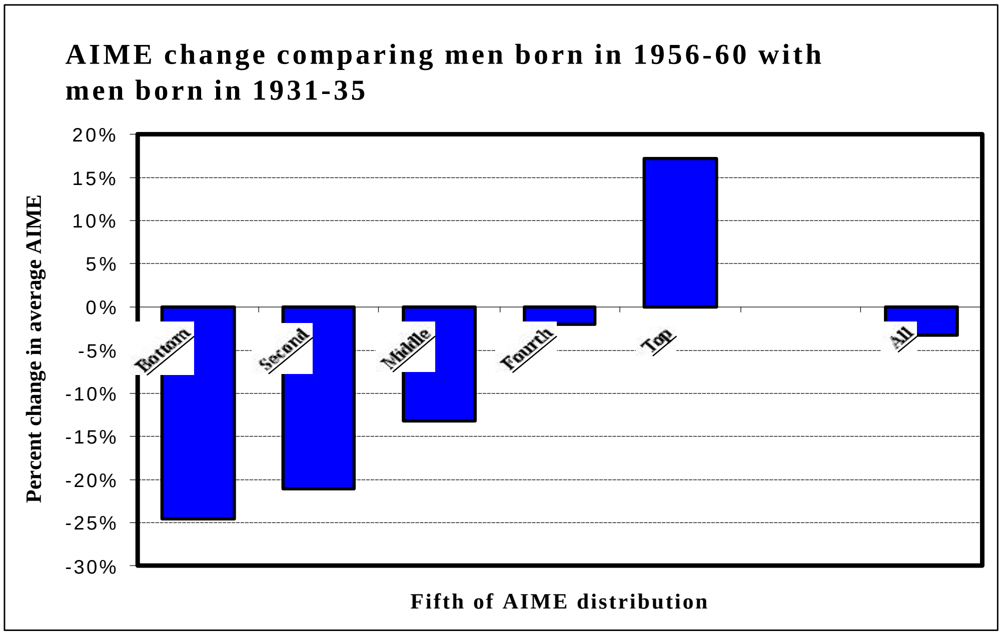


The AIME distribution has grown more unequal over time both for men and women, though the pattern differs somewhat across the two sexes. Men at the top of the earnings distribution have experienced an accelerating rise in the proportional distance between their earnings and those of men in the middle. Men at the bottom suffer only a small decline in their relative earnings compared with men in the middle fifth. In contrast, among women the upward trend in relative earnings at the top of the distribution is more moderate, but the downward drift of relative earnings at the bottom is faster than it is among men. One explanation is that women in the top four fifths experience *gains* in their average AIMEs. This is partly because women in more recent cohorts are more steadily employed throughout their careers, and hence have fewer years with zero earnings than women in earlier cohorts. It is also partly due to the fact that, in comparison with more recent cohorts of men, more recent cohorts of women have experienced faster gains in both hourly earnings and hours worked if they are employed. Figure 2-9 shows the percentage change in average AIME, within each fifth of the AIME distribution, if we compare the more recent with the earliest cohorts in the sample. The right-hand side of the figures shows the *average* change among workers in all parts of the distribution when comparing workers born in 1956-1960 with those born in 1931-1935. The average AIME of women born in 1956-1960 is 30 percent higher than the average AIME of women born between 1931 and 1935. In contrast, the average AIME of men born in 1956-1960 is 3 percent *lower* than the average AIME of men born between 1931 and 1935. Both among men and women the AIME gains are fastest among workers in the top fifth of the AIME distribution. But in contrast to the poor performance of the AIME in the middle three fifths of the male AIME distribution, women in the same positions in the female AIME distribution have experienced increases in their earnings relative to economy-wide average earnings.

In sum, these estimates show that women have made and will continue to make earnings gains compared with men. Workers of either sex will also experience substantial increases in lifetime earnings inequality, mirroring the annual pattern of growing earned income inequality the nation has witnessed over the past twenty years. Finally, workers born in the middle and toward the end of the Baby Boom will experience smaller lifetime earnings gains compared with the first Baby Boom cohort. Workers born immediately after World War II had significantly higher educational attainments than the generations born before them, but successive cohorts of Baby Boomers did not sustain the rapid gains in schooling attainment that earlier generations achieved. The later Baby Boom cohorts also faced the misfortune of entering the labor force when the relative earnings of young workers fell. Indeed, for men in these cohorts *absolute* as well as relative earnings declined. The bad fortune will leave typical members of the later Baby Boom with lower *relative* career wages than those earned by the first cohort born after the Second World War.

Figure 2-9
Change in Average AIME
Within Fifths of AIME Distribution, by Sex and Birth Cohort





Problems with the projections.

There are two main problems with the predictions generated using our method. The more serious problem is that while the procedure generates predictions that have approximately correct expectations and error distributions for most people, the method probably does not generate accurate predictions of the number of years with positive earnings. We are also skeptical that the estimates of earnings can be used to generate wholly reliable estimates of the number of quarters of covered earnings that workers earn in given years. It would probably be preferable to estimate a separate equation to predict how many quarters of coverage a worker obtains for a given expected value of covered earnings in a year. Ultimately, of course, it is desirable to estimate a lifetime earnings model that produces useful predictions of the age of retirement or permanent disability as well as the pattern of pre-retirement annual earnings.

A less serious problem is that the procedure appears to generate excessive predictions of earnings and employment for persons reaching their sixties between 1997 and 2010. The problem seems to be caused by our failure to take proper account of the effect of error truncation in making predictions. The problem is much more serious when expected earnings are very low (as they are when workers reach their sixties) than when expected earnings are higher. The problem will have only a small effect on estimates of the AIME at age 62, but the effect on estimates of the AIME at age 67 will be more noticeable, especially among women. As noted earlier, however, subsequent chapters in this report do not utilize all the estimates in this chapter of earnings between ages 62 and 67. Instead, in Chapters 5 and 6 of this report, we present separate models for workers aged 62 to 67 which forecast the initial year of acceptance of Social Security benefits and then forecast "partial retirement" earnings for those workers who have chosen to receive benefits. The projections in this chapter are used only to forecast the earnings of those workers over aged 62 who do not receive retirement or disability benefits.

APPENDIX A

PROCEDURE FOR ESTIMATING EARNINGS FOR UNOBSERVED FORMER HUSBANDS OF DIVORCED WOMEN

I. INTRODUCTION

An individual's Social Security benefit is the highest among those determined by 1) their own work, 2) their current spouse's work, or 3) under some circumstances, the work history of a divorced or deceased spouse. Because of these alternative ways of computing benefits, it is necessary to ascertain earnings records for multiple spouses to determine the individual's Social Security benefit. The projection file from Rand provides spouse identifiers for spouses observed at the time of the SIPP interview, but no spouse identifiers for marriages that ended before the SIPP panel started or marriages that are projected to occur after the SIPP panel ended. The projection sample does not reveal the earnings and qualification history for former or future spouses. In order to determine Social Security entitlement, we must impute spouses for each unobserved marriage.

Over half of all spouses are unobserved. The Rand projection file has 113,071 individuals comprising 149,445 marriages. About 6 percent (7,302 individuals) of individuals never get married. For about 42 percent (62,102 marriages) of marriages, we observe the real spouse. All other spouses get imputed (87,343 marriages).

The initial RFTOP only required estimating earnings for unobserved former husbands of divorced women. The problem of unobserved spousal earnings extends beyond just divorced women. It also extends to new spouses (i.e., marriages that occur after the SIPP panel) and deceased spouses. On the projection sample, about 18 percent of women get married after the last interview on the SIPP, about 9 percent of women in unobserved marriages get divorced after 10 years of marriage, and about 6 percent of women in unobserved marriages are widowed before age 60.

We used a statistical matching algorithm to find a spouse from the projection sample with the characteristics specified in the demographic projections from Rand (Part II of the TO). For observed marriages, we match to the actual spouse. For unobserved marriages, we impute a spouse. After matching spouses, Social Security entitlement for any individual for any spouse is simply a matter of looking at the marriage and work characteristics for each spouse to see if he or she meets the Social Security entitlement rules. This match works for current spouses, divorced spouses, and deceased spouses.

II. BACKGROUND

Rand's projection file contains demographic and marriage information for each individual born between 1926 and 1965. The demographic information includes the individual's date of birth, race, educational attainment, hispanicity, a measure of permanent income, disability status, date of onset of disability, and date of death. The marriage information includes, for up to 9 marriages, the marriage starting date, the marriage ending date, and the marriage termination status. In addition, it includes the demographic information for each spouse (except for age of death).

Where possible, Rand assigned this information from the SIPP core data and the marriage history topical module. When this information was not available, such as all information for marriage and mortality characteristics that occur after the SIPP panel ended and the demographic information for all marriages that ended before the SIPP panel began, Rand projected it. This file gives us a great deal of information about the characteristics of each spouse. For each person for each marriage, we find the individual on the projection file that best matches Rand's spousal characteristics. This person provides the missing earnings and work history characteristics necessary to calculate individuals' Social Security benefits.

III. METHODOLOGY

To impute spouses, we used a statistical matching algorithm based on minimizing a distance function. We limited the pool of potential spouses to those individuals of the proper gender born within two years of the desired birth year. Within the set of potential donors, we selected the "best" individual to be the spouse, where "best" is defined to be the individual with the smallest distance measured by a distance function. The characteristics in the distance function are limited to those that Rand projects: spouse's birth date, marriage start date, marriage end date, marriage termination status (divorce, widow, death), spouse's disability status, disability date, race, hispanicity, education, and permanent income.

The distance function is defined as follows:

$$D_d = \sum_{j=1}^{n} w_j*[(X_{dj} - X_{rj})/\sigma_j]^2$$

where j is the number of measured attributes in the distance function, w is a weight factor, X is a characteristic measure (permanent income, marriage start date, race, educational attainment, etc.), σ is the standard deviation of the jth X variable in the dataset, d denotes the characteristic of the donor, and r denotes the characteristic of the recipient. The statistical match finds the best match among all records in the donor pool.

The weight factor, w_j, allows the analyst to decide which attributes are more important to match on. For example, if it is critical to get the race and educational attainment of the donor right, these attributes should get relatively higher weights. If race is more important than educational attainment, it should have a higher weight. We calculated the distance, D, for each donor record, and selected the donor record with the smallest value to be the spouse.[5]

We assigned weight factors to optimize the earnings match. Because the purpose of the imputation is essentially to impute a summary earnings record for each spouse, we selected larger weights on characteristics likely to affect earnings. These include permanent income, date of death, and date of disability. Less important characteristics received lower weights. These include marriage start date, marriage termination status, race, and education. The standard deviation in the distance function scales the differences to a common unit and reduces the impact of highly variable characteristics.[6] These values are based on the standard deviation of the spouse variables for first marriages. The specific values used in the distance function are shown in Table 2-A-1.

Table 2-A-1
Weights and Standard Deviations Used in the Distance Function

Variable	Weight	Standard Deviation
Birth date	3	4320.99
Hispanicity	1	0.2761
Education	1	0.6358
Race	1	0.6011
Death date	2	6580.35
Disability date	2	6258.08
Disability status	1	0.2973
Permanent income	5	0.7632
Marriage start date	1	5170.19
Marriage end date	5	8431.54
Marriage termination status	1	0.8090

Source: The Urban Institute.

Date of death is a very important criterion in selecting a potential spouse. If a potential spouse dies before the marriage ends, his or her earnings would be incorrectly censored at the individual's date of death. Rand did not project the date of death for spouses. Instead, they projected the date of marriage termination and the termination status. When the termination status is widowhood, we know the date of death. When the termination status is divorce, we know only that the individual survived until the marriage termination date. In selecting spouses, we only consider individuals who survive at least to age 70 (after the completion of earnings for most individuals) or the marriage termination date.

Similarly, the date of onset of disability (work disability as opposed to Social Security disability) is another important criterion in selecting a potential spouse. If a potential spouse becomes unable to work due to a disability, his or her earnings would be censored or reduced at the individual's disability date. For individuals who do not become disabled, Rand did not assign a disability date. To prevent a missing value from entering the distance function, we assigned disability date to the date of death for individuals not projected to become disabled.

Individuals married at the time of the SIPP panel always select their actual spouse in the marriage match. Rand did joint projections when they assigned marriage characteristics and mortality. Thus, husbands and wives marriage characteristics are internally consistent.

IV. RESULTS

The quality of the match can be evaluated by comparing what we wanted with what we received. A perfect match is one for which the donor record matched each specific criteria we matched against. For example, if we wanted a black, college graduate, born in 1950, we would say we made a good match if we found a black, college graduate, born in 1950. Because many of the characteristics in the distance function are continuous variables, it is extremely unlikely that an exact match exists within the donor pool.

The projection data set consists only of individuals born between 1926 and 1965. Projected spouses, however, are not constrained to be from these birth cohorts. Because the individuals of interest are those born between 1931 and 1960, most spouses will be within the projection dataset. The exceptions are those individuals on the tails of the cohort distribution with considerably older or younger spouses. The best match for these individuals will come from individuals in the tails. For example, if we want someone born in 1920, we match to someone born in 1926 (the earliest birth cohort in our sample). If we want someone born in 1970, we match to someone born in 1965 (the latest birth cohort in our sample). Earnings in every year are systematically off by the difference in the number of years between the age of the imputed and projected spouse. We align the dates of earnings on the imputed spouse to match the age of the projected spouse. In the latter case, for example, the earnings on the matched record (imputed spouse) in 1985 is for a 20 year old, but the spouse I want (projected spouse) is only 15 years old. After the alignment, age 20 (year 1985) earnings from the imputed spouse is age 20 (year 1990) earnings on the projected spouse. We make this adjustment for every year of earnings. For our projection sample (born between 1931 and 1960 with a full panel weight) 1.6 percent are married to someone born before 1926 and 0.6 percent are married to someone born after 1965.

On basic demographic characteristics (race, hispanicity, education, birth date), the match works very well. For real spouses, these characteristics are observed, and we get a perfect match. For the imputed spouses, we find an exact match in almost 100 percent of marriages on hispanicity, 93 percent on educational attainment, and 96 percent on race (see Table 2-A-2). For 91 percent of imputed marriages, we are within two years of the desired date of birth. The donor pool was limited to individuals within two years of birth. All births outside this range are for marriages to individuals born before 1926 or after 1965.

Table 2-A-2
Percent of Imputed Spouses Who Match Specific Characteristics

Characteristic	Imputed Spouse
Birth date (within 2 years)	90.6%
Hispanicity	99.9%
Education	92.6%
Race	96.3%
Death date (within 3 year)	83.2%
Disability date of those projected to become disabled (within 3 years)	36.4%
Disability status	99.6%
Permanent income (within 0.3)	87.9%
Marriage start date (within 3 years)	39.0%
Marriage end date (within 3 years)	47.4%
Marriage termination status	91.8%

Source: The Urban Institute.

Match quality on mortality is comparatively poor. Eighty-three percent of imputed spouses die within three years of the date of death. This statistic is misleading, however, because we really only care about mismatched date of death when the individual dies before the earlier of retirement age (62 or 67) or end of marriage date. For example, if we want a spouse who dies at age 80 and we find a spouse who dies at age 70, the match statistic would indicate that the date difference is 10 years. The earnings history, however, for the spouse would, for the most part, be complete. We limited donors to those individuals who die after age 69 or no sooner than one year before the marriage termination date.

For almost all cases (99.6 percent) where Rand projects spousal disability (disability limits work), we match to someone projected to become disabled. Thirty-six percent of these imputed spouses become disabled within three years of the projected date of onset, and 60 percent become disabled within five years of the projected date of onset. The probability of becoming disabled is lower at younger ages. As with mortality, the match quality on date of disability is comparatively poor for young disability because of small sample size.

We match very well on permanent income. Permanent income is a measure of the family income relative to the average wage. We heavily weighted this characteristic in hopes of capturing potential earnings of the matched spouse. For 88 percent of imputed spouses, the difference in permanent income was less than 0.3.[7]

Earnings are less likely to be associated with marriage start and end dates than with educational attainment and permanent income. Therefore, we gave marriage start and end dates relatively low weights. As such, the match quality of these variables is comparatively poor. For imputed spouses, only 39 percent of marriage start dates are within 3 years and only 47 percent of marriage end dates are within 3 years. This mismatch is unlikely to affect men’s earnings. It might affect the timing and duration of drop-out years due to women’s child bearing, though these are largely correlated with age and education on which we match quite well.

Marriage termination status is really a measure of mortality. Marriages either end in divorce or death. The quality of the match on termination status is a direct reflection of the quality of the mortality match. As with mortality, we do well on termination status for likely outcomes (death in late years) and poorly on unlikely outcomes (death in early years). For example, the donor pool of individuals who get divorced at young ages is big compared to the donor pool of individuals who die at young ages. If we want someone who divorces at a young age, we match well. If we want someone who dies young, we match comparatively less well. For 92 percent of imputed spouses, we match the marriage termination status.

V. FILE STRUCTURE

The spouse match program creates a file that can easily be used to access any spousal characteristic, including earnings, for multiple marriages. This file contains a random access pointer for each marriage (spindex1-spindex9), a spouse identifier for each marriage equal to PPID*10,000,000 + PPENT*10,000 + PPNUM*10 + PANEL (spid1-spid9) where PANEL ranges from 0 to 3 representing 1990 to 1993, and a spouse imputation flag for each marriage (simpute1-simpute9) where 0=not imputed and 1=imputed. In addition, the file has SIPP and ID for matching to other files.

The strength of the random access spouse pointer is that it does not require multiple sorting and merging. For any spouse of interest, you can use the point option on the SAS set statement to merge on the desired spousal characteristics. To do this requires two things: first, the file you point to must be sorted by SIPP and ID and have 113,071 observations, and second, every spousal characteristic of interest must be explicitly kept and renamed. The first requirement is due to the fact that the pointer to the spouse's record number is based on the full universe of records sorted as specified. The second requirement is due to the fact that any variable on the spouse file that shares a name with the base file will take on the spouse's value. After renaming the variable, the merged file has both the base characteristic and the spouse characteristic.

There is nothing in the data structure of the spouse file that precludes sorting and merging spouses in the traditional fashion. In the case of a woman, for example, with Social Security entitlement to three deceased spouses' Social Security, the sort and merge method becomes cumbersome. Random access pointers allow you to check each spouse's characteristic in one data pass. This allows for easier data management and simpler data processing.

VI. CONCLUSIONS

The quality of the spouse match seems quite high based on characteristic by characteristic comparisons. The match is particularly good for the basic demographic characteristics (hispanicity, education, race). The match is worse in areas outside of our control, such as the limited birth cohort inclusion. After adjusting the earnings series for these cases, this limitation is no longer problematic.

The use of random access spouse pointers simplifies the task of assessing spousal characteristics for multiple marriages. Random access pointers give the analyst access not only to spousal earnings, but also to pensions, wealth, Social Security participation, and partial retirement earnings. This data structure provides a flexible and powerful connection to spousal characteristics.

CHAPTER 2: REFERENCES

Burtless, G., 1995. "International Trade and the Rise in Earnings Inequality." *Journal of Economic Literature 33* (2) 800-16.

Freeman, R. B., 1997. *When Earnings Diverge: Causes, Consequences, and Cures for the New Inequality in the U.S.* Washington, DC, National Planning Association.

Iams, H.M., and Sandell, S.H., 1997. "Projecting Social Security Earnings: Past Is Prologue." *Social Security Bulletin 60* (2) 3-16.

Levy, F., and Murnane, R.J., 1992. "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature* 30 (3) 1333-81.

O'Hare, John, 1997. "Impute or Match? Strategies for Microsimulation Modeling," A paper presented at Microsimulation in Government Policy and Forecasting International Conference on Combinatorics, Information Theory and Statistics. 1997.

CHAPTER 2: LIST OF TABLES

Table 2-1
Female Age-Earnings Profiles, by Educational Attainment
Fixed-Effect Model Estimates 17
Table 2-2
Male Age-Earnings Profiles, by Educational Attainment
Fixed-Effect Model Estimates 19
Table 2-3
Probit Model of Disability Insurance Onset
by Gender, 1987 - 1994 24
Table 2-A-1
Weights and Standard Deviations Used in the Distance Function 42
Table 2-A-2
Percent of Imputed Spouses Who Match Specific Characteristics 44

CHAPTER 2: LIST OF FIGURES

Figure 2-1
Age-Earnings Profile of U.S. Women,
Including and Excluding Zero Earnings 9
Figure 2-2
Age-Earnings Profile of U.S. Men,
Including and Excluding Zero Earnings 9

Figure 2-3
Estimated Age-Earnings Profiles
By Sex and Educational Attainment 21
Figure 2-4
Trend in AIME by Birth-Year Cohort: Both Sexes 31
Figure 2-5
Trend in AIME by Birth-Year Cohort: Both Sexes 32
Figure 2-6
Distribution of Educational Attainment by Birth Cohort 34
Figure 2-7
Trend in AIME by Birth Cohort and Fifths of AIME Distribution:
Both Sexes 35
Figure 2-8
Trend in Relative AIME by Birth Cohort
Within Fifths of the AIME Distribution 36
Figure 2-9
Change in Average AIME
Within Fifths of AIME Distribution, by Sex and Birth Cohort 38

CHAPTER 2: ENDNOTES

1. The out-of-sample projections described below pertain to the sample members born between 1931 and 1960, since these people were the principal focus of the study. The estimates were derived using a sample that included people born between 1926 and 1965 to improve the estimation of the earnings function at older ages and to generate earnings predictions for people outside the 1931-1960 frame. In other parts of this project, these estimates are needed to estimate the distribution of earnings among people who might marry or divorce people born between 1931 and 1960.

2. The forecasts of Average Indexed Monthly Earnings, discussed below, are based on earnings reports and earnings predictions *below* the maximum taxable wage, that is, on the covered wages that are actually used by the Social Security Administration to calculate benefits.

3. The age-earnings profiles of college graduates and workers with post-college education have a somewhat different pattern (earnings of people with advanced degrees are sharply lower at early ages, for example), but the two profiles seem to have a similar average level. This is

misleading. The average value of the individual-specific effect probably differs for workers with college and post-graduate degrees, implying that the average level of earnings – not just the pattern of rise and fall over time – also differs between the two groups.

4. The actual AIME of a worker who is predicted to receive a DI pension is calculated at the age of predicted DI onset. This nominal earnings estimate is then indexed through the calendar year that the worker attains 62 and is compared with economy-wide average earnings at age 62. Thus our estimate of the AIME for both DI and OAI beneficiaries is calculated relative to economy-wide earnings in the same year, the year the worker reaches age 62.

5. For a description of statistical matching techniques, see O'Hare (1997).

6. All dates on the file are stored as number of days since January 1, 1960.

7. Permanent income has a mean of 0.127 and a standard error of 0.750. We chose the critical value for permanent income to be within a 90 percent confidence interval from a two-tailed normal distribution.